Федеральное государственное автономное образовательное учреждение
высшего образования «Российский университет дружбы народов»

На правах рукописи

Белов Александр Александрович

Обобщение метода конечных разностей на задачи с особенностями в решении

Специальность 1.2.2. Математическое моделирование, численные методы и комплексы программ

Диссертация на соискание ученой степени
доктора физико-математических наук

Научный консультант
доктор физико-математических наук, профессор
Севастьянов Леонид Антонович

Москва—2023

Содержание

Введение

Актуальность темы исследования

В физике и технике возникают новые все более сложные задачи, предъявляющие чрезвычайно требования к точности и надежности расчета. Это приводит к бурному развитию приближенных методов расчета, из которых наиболее универсальными являются методы конечных разностей (МКР) и конечных элементов (МКЭ). В рамках этих подходов разработано большое количество алгоритмов как общего назначения, так и ориентированных на конкретные прикладные задачи. Эти алгоритмы реализованы в широко известных прикладных пакетах.

Применение традиционных алгоритмов МКР к новым задачам сталкивается со значительными трудностями: потеря точности, замедление или отсутствие сходимости, потеря устойчивости и т.д. В этом случае обобщение МКР на новые классы задач требует разработки новых алгоритмов, повышающих точность и надежность этого метода.

Большой вклад в развитие МКР был сделан чл.-корр. РАН Н.Н. Калиткиным и его учениками: предложенные ими алгоритмы позволили значительно расширить круг задач, которые удается успешно решать с помощью МКР.

Алгоритмы, развиваемые школой Калиткина, имеют следующие особенности. Во-первых, они включают многосеточный расчет с апостериорным контролем фактической погрешности и реккурентным повышением точности. Основы этого подхода были заложены в работах Ричардсона. Во-вторых, для краевых задач используются консервативные схемы. Калиткиным были впервые введены бикомпактные схемы, обобщающие консервативные схемы Самарского. В-третьих, для начальных задач широко применяется параметризации через длину дуги интегральной кривой.

Однако, несмотря на достигнутые успехи, ряд задач по-прежнему не удается решить в рамках существующих реализаций МКР. ***Общим свойством*** этих задач является наличие принципиальных трудностей, связанных с наличием

особенностей в решении: пограничных слоев (которые в пределе при уменьшении ширины стремятся к разрывам), сингулярностей (в которых решение или его производные обращаются в бесконечность), разрывов на границах раздела сред. Поэтому разработка новых подходов, позволяющих решать данные задачи в рамках МКР, является ***актуальной***.

Разработанность темы диссертации

1) До сих пор значительные трудности представляет расчет жестких и плохо обусловленных задач Коши (к ним относятся кинетика реакций, модели нелинейных осцилляторов, системы ОДУ, полученные в результате применения метода прямых к УрЧП и др.). Значительный вклад в развитие соответствующих численных методов внесли работы следующих авторов. Дж. Гиршфельдер и Ч. Кертисс предложили одну из первых численных схем для жестких задач. Дж. Бутчер систематически развивал теорию схем Рунге-Кутты и впервые построил большое количество схем высоких порядков точности. В работах Х. Розенброка были предложены исключительно удачные явно-неявные схемы. Дальнейшему их развитию посвящен работы Э. Хайрера, Г. Ваннера, С.С. Филиппова, Е.А. Новикова. Особенно отметим работы представителей школы Калиткина: П.Д. Ширков, А.Б. Альшин и Е.А. Альшина впервые предложили двухстадийные явно-неявные схемы с комплексными коэффициентами. В работах Калиткина и И.П. Пошивайло предложены оптимальные обратные схемы Рунге-Кутты, обладающие одновременно высокой точностью и уникальной надежностью.

Для расчетов жестких задач широко используют алгоритмы автоматического выбора шага, основанные на локальном сгущении сеток или вложенных схемах. Наиболее известными являются алгоритмы Ч. Гира и Дж. Дормана, П. Принса. Дальнейшему развитию этих подходов посвящены работы Бутчера, Л. Шампина, П. Капса, Дж. Кэша, Дж. Прентиса и ряда других исследователей. Однако вопрос о фактической точности таких расчетов остается открытым. В

литературе неоднократно отмечалось, что фактическая точность порой не соответствует заданной пользователем. Известны также случаи, когда эти алгоритмы не позволяют завершить расчет, срываясь до того, как достигнут конец отрезка интегрирования.

Характерным примером жесткой задачи является задача кинетики химических реакций. Эта задача имеет ряд специфических особенностей; в частности, ее решение является неотрицательным. Калиткин и В.Я. Гольдин построили специализированную разностную схему для этой задачи, обеспечивающую неотрицательность численного решения. Эта схема была очень надежной, однако она имела первый порядок точности.

2) Не меньшие трудности представляет исследование решений, имеющих подвижные особые точки, что типично для нелинейных моделей. Примерами являются нелинейное горение, кумуляция, пробой в плазме и т.д. В литературе описано большое количество подходов. М. Бергер и Р. Кон предложили метод масштабирования. Развитием этого подхода является применение нормализующих групп, предложенное в работах Л. Чена и соавторов. В работах А. Кангиани и соавторов были предложены адаптивные сетки на основе априорных и апостериорных оценок решения. Ч. Бадд и Р. Рассел предложили и программно реализовали метод движущихся сеток. В работах П. Гройсмана для задач с сингулярностями предложено использовать чисто неявные схемы (например, обратную схему Эйлера). В работах Ч. Чо введено понятие численного разрушения и предложен алгоритм его обнаружения.

Наиболее работоспособным является метод Альшиной-Калиткина-Корякина, основанный на анализе сходимости при сгущении сеток. Он позволяет рассчитать положение и порядок особенности, но вопрос о фактической точности такого расчета остается открытым.

3) В ряде приложений (например, прямое вычисление некоторых специальных функций) требуется расчет решения ОДУ не только до полюса, но и после него. В работах Б. Форнберга, Дж. Вайдемана, А.А. Абрамова, Л.Ф. Юхно

разработан ряд подходов для расчета трансцендент Пенлеве. В работах М.Д. Малых и Л.А. Севастьянова был разработан способ продолжения за полюс для уравнений с квадратичной нелинейностью. Однако разностные методы, единообразно применимые к широкому классу задач, не предложены.

4) В ряде актуальных проблем требуется решать систему уравнений Максвелла, описывающих эволюцию электромагнитного поля в слоистых средах. В стационарном случае для таких задач наиболее популярны следующие методы. Современная формулировка метода матриц рассеяния построена Д. Берреманом; систематическое применение этого метода к прикладным задачам представлено в работах А.Г. Свешникова, А.В. Тихонравова, Севастьянова, К.П. Ловецкого, А.А. Хохлова. Методы типа RCWA предложены в работах М. Мохарама и Т. Гейлорда. Этот класс методов активно развивается в работах Севастьянова, Ловецкого, Хохлова, Егорова, В.А. Сойфера. Основополагающие результаты по методу конечных элементов получены в работах Неделека, Равьяра и Томаса, Хистхэвена; систематическому применению этого метода посвящены работы В.В. Котляра и Сойфера. Современная формулировка метода конечных разностей в частотной области приведена в монографии Ю. Инана и Р. Маршалла. В нестационарном случае наиболее употребительными являются различные варианты метода конечных разностей во временной области. Большой вклад в развитие этого подхода внесли Инан, А. Тафлов, Д. Салливан, Дж. Беренгер, Г. Мур, Сойфер, Котляр, Свешников, А.Н. Боголюбов. В литературе отмечено, что для сеточных методов значительную проблему представляет потеря точности вблизи границ раздела и учет частотной дисперсии среды.

Цель диссертационной работы

Целью диссертационной работы является обобщение метода конечных разностей на задачи с вычислительными особенностями: контрастные структуры, подвижные особые точки, разрывы коэффициентов.

С этой целью были рассмотрены математические модели, имеющие указанные особенности: жесткие задачи Коши для ОДУ (включая задачу кинетики реакций), задачи Коши с сингулярностями в решении (включая уравнения в частных производных, например, модели нелинейного горения), задачи для системы одномерных уравнений Максвелла в слоистых средах с частотной дисперсией.

Задачи диссертационной работы

Для достижения указанной цели решены следующие задачи:

1. Разработка, обоснование и тестирование алгоритма выбора шага по кривизне интегральной кривой для численного интегрирования задач Коши для ОДУ.
2. Разработка, программная реализация и тестирование явной схемы для расчета кинетики реакций, которая имеет второй порядок точности и обеспечивает неотрицательность численного решения. Сравнение предложенной схемы с другими методами первого и второго порядка точности.
3. Разработка, обоснование и тестирование методов численного исследования подвижных особых точек и их последовательностей в решениях ОДУ с апостериорной асимптотически точной оценкой погрешности. Реализация предложенных методов в виде комплекса проблемно-ориентированных программ.
4. Применение предложенных методов к исследованию сингулярностей в решениях УрЧП методом прямых.
5. Разработка, обоснование, программная реализация и тестирование бикомпактной разностной схемы для одномерной системы уравнений Максвелла в слоистых средах. Сравнение предложенных схем с другими методами (МКЭ, МКР во временной области).

6. Разработка, обоснование и тестирование метода интегрирования уравнений Максвелла вдоль оптического луча для задачи о наклонном падении плоской волны на набор плоско-параллельных пластин. Реализация предложенного метода в виде комплекса проблемно-ориентированных программ.

Научная новизна

Для численного интегрирования задачи Коши для ОДУ предложен, реализован и протестирован новый метод построения квазиравномерных сеток – геометрически-адаптивных сеток – на основе оригинального подхода к выбору шага по кривизне интегральной кривой.

Предложена, реализована и протестирована новая специальная явная схема второго порядка точности для расчетов кинетики реакций, обеспечивающая положительность решения.

Оценены границы применимости сеточных методов и методов разложения по малому параметру, и дан ряд практических рекомендаций.

Предложен, реализован и протестирован новый метод обнаружения и исследования алгебраических особых точек и логарифмических полюсов для систем ОДУ, который позволяет рассчитывать параметры особенности с апостериорной асимптотически точной оценкой погрешности.

Предложен, реализован и протестирован новый метод обобщенной инверсной функции для решения задачи Коши для систем ОДУ с последовательностью алгебраических особых точек целого порядка. В отличие от ранее известных подходов, предложенный метод не использует априорной информации о свойствах задачи.

Для системы стационарных и нестационарных одномерных уравнений Максвелла предложена бикомпактная консервативная разностная схема. Для предложенных схем доказана сходимость для произвольных неравномерных сеток и неоднородных сред.

Для задачи о наклонном падении плоской волны на систему плоско-параллельных либо клиновидных пластин предложен метод интегрирования уравнений Максвелла вдоль оптического луча, позволяющий решать эту задачу по одномерным схемам.

Теоретическая и практическая значимость работы

Предложенные математические методы качественно превосходят по точности, надежности и экономичности ранее известные алгоритмы, расширяют область приложения метода конечных разностей и представляют интерес для широкого круга исследователей при решении прикладных задач. Они используются в работах научной группы проф. Малых в Российском университете дружбы народов, могут непосредственно использоваться в работах, проводимых на ряде факультетов МГУ им. М.В. Ломоносова (физическом, химическом, механико--математическом, ВМиК), в НИИ Механики МГУ, в Институте проблем механики им. А. Ю. Ишлинского РАН, в Институте прикладной математики им. М. В. Келдыша РАН, в федеральных ядерных центрах – ВНИИЭФ (Саров) и ВНИИТФ (Снежинск), в Физическом институте академии наук им. П. Н. Лебедева, в Объединенном институте ядерных исследований, в НИЦ "Курчатовский институт" и ряде других организаций.

Геометрически-адаптивные сетки следует рассматривать как надежный метод расчета жестких задач Коши для ОДУ; они должны заменить существующие программы, основанные на вложенных схемах и локальном сгущении сетки.

Методы диагностики сингулярностей и расчета задач со множественными полюсами могут стать надежным инструментом для исследования задач с разрушением решения, проводимых на кафедре математики Физического факультета МГУ им. М. В. Ломоносова в группе проф. М.О. Корпусова, в Российском университете дружбы народов в группе проф. Л.А. Севастьянова, в Московском

институте электронной техники в группе проф. Г.Л. Алфимова и в других организациях.

Построенные методы решения уравнений Максвелла применимы к различным задачам электродинамики слоистых сред (фотоника, плазмоника) и могут быть использованы в СГУ группой проф. В.Л. Дербова, в РУДН в группе проф. Севастьянова, на физическом факультете МГУ в группах проф. А.Н. Боголюбова и проф. А.А. Федянина, на факультете ВМиК МГУ в группе проф. Ю.А. Еремина, в НИУ ИТМО на мегафакультете фотоники под руководством проф. П.А. Белова и ряде других организаций. Предложенные методы могут стать надежным вычислительным инструментом в расчетах, предваряющих натурный эксперимент. Выполненные расчеты динамики поверхностных волн Блоха могут непосредственно использоваться при планировании новых экспериментов, которые проводятся в указанных коллективах на физическом факультете МГУ и в НИУ ИТМО.

Методы исследования

При разработке математических алгоритмов использовались традиционные методы вычислительной математики. Разностные схемы составлялись методом разностной аппроксимации. Для тестирования предложенных методов проводились расчеты задач с известным точным решением. Вычислялась погрешность численного решения относительно точного. Дополнительно вычислялись оценки точности по Ричардсону.

При разработке прикладных пакетов был использован язык Matlab, совместимый со свободно распространяемой средой для математических вычислений GNU Octave. Она позволяет легко визуализировать результаты расчетов.

Положения, выносимые на защиту

На защиту выносятся следующие положения:

1. Для жестких задач Коши предложен, программно реализован и протестирован метод геометрически-адаптивного выбора шага по кривизне интегральной кривой. На его основе проведено оригинальное исследование кинетики реакций водород-кислородного горения.
2. Для задач кинетики реакций предложена и протестирована явная схема второго порядка точности, обеспечивающая неотрицательность численного решения.
3. Предложены и протестированы методы численного интегрирования ОДУ с вещественными подвижными сингулярностями. Они позволяют численного обнаружить и исследовать ближайшую особую точку алгебраического и логарифмического типа, а также проводить расчет решения с последовательностью алгебраических особых точек целого порядка.
4. Для системы одномерных уравнений Максвелла в слоистой среде с частотной дисперсией предложена и протестирована бикомпактная разностная схема.
5. Предложен и протестирован метод оптических путей для интегрирования уравнений Максвелла вдоль оптического луча, позволяющий рассчитывать ряд двумерных задач с помощью одномерных схем.

Достоверность и обоснованность полученных результатов

Обоснованность полученных результатов обусловлена тем, что для всех предложенных методов доказаны теоремы о сходимости.

Предложенные методы проверялись на представительных тестовых задачах с известным точным решением. В ходе расчетов непосредственно проверялась

сходимость численного решения к точному, а также контролировалось соответствие фактического порядка точности теоретическому.

Расчеты прикладных задач, точное решение которых неизвестно, проводились на сгущающихся сетках с апостериорной оценкой погрешности по методу Ричардсона и контролем фактического порядка точности. Это обеспечивает математическую точность на уровне ошибок округления компьютера.

Достоверность полученных результатов обеспечивается тем, что расчеты по предложенным алгоритмам сравнивались с расчетами по другим широко известным методам либо с доступными результатами натурного эксперимента.

Апробация результатов

Результаты работы докладывались на следующих конференциях: международные конференции «Современные проблемы вычислительной математики и математической физики» памяти академика А.А. Самарского к 95-летию со дня рождения и к 100-летию со дня рождения (Москва, июнь 2014, июнь 2019), международная конференция «Современные проблемы математической физики и вычислительной математики» к 110-летию со дня рождения академика А.Н. Тихонова (Москва, ноябрь 2016), XVII, XIX, XXI международные конференции «Харитоновские тематические научные чтения» (Саров, март 2015, апрель 2017, апрель 2019), международная конференция Progress In Electromagnetics Research Symposium (Санкт-Петербург, май 2017), международная конференция 13th annual workshop «Numerical methods for problems with layer phenomena» (Москва, апрель 2016), Научно-координационная сессия по неидеальной плазме (Москва, ноябрь 2015), XVI, XVII Всероссийские школы-семинары «Физика и применение микроволн» им. проф. А.П. Сухорукова (Можайск, июнь 2017, июнь 2019, июнь 2021), XVII, XVII Всероссийские школы-семинары «Волновые явления в неоднородных средах» им. проф. А.П. Сухорукова (Можайск, июнь 2018, июнь 2020) научная конференция «Ломоносовские чтения» (Москва, ап-

рель 2016, апрель 2021), конференция Совета молодых ученых ИПМ им. М. В. Келдыша РАН (Москва, ноябрь 2015), конференция, посвященная 90-летию со дня рождения В.Я. Гольдина (ИПМ им М.В. Келдыша РАН).

Сделаны доклады на следующих научных семинарах: семинар кафедры вычислительной математики ВМиК МГУ им. М.В. Ломоносова (декабрь 2013), семинары кафедры математики физического факультета МГУ им. М.В. Ломоносова (март 2016, декабрь 2016), семинар по обратным задачам математической физики НИВЦ МГУ им. М.В. Ломоносова (ноябрь 2016), семинар отдела № 14 ИПМ им. М.В. Келдыша РАН (январь 2017), семинар Института прикладной математики и телекоммуникаций РУДН (Москва, апрель 2021, май 2022), семинар по вычислительной и прикладной математике Лаборатории информационных технологий им. М.Г. Мещерякова ОИЯИ (Дубна, ноябрь 2021).

Личный вклад автора

Все результаты диссертации, выносимые на защиту, получены автором лично. В работах, опубликованных в соавторстве, вклад автора является определяющим.

Соответствие паспорту специальности

В диссертации разработаны новые математические методы моделирования объектов и явлений: экономичные методы моделирования задач кинетики реакций, процессов нелинейного горения, задач интегральной фотоники и ряда других (п. 1 паспорта специальности 1.2.2).

Разработаны, обоснованы и протестированы новые эффективные вычислительные методы для задач с особенностями в решении – жестких задач Коши для ОДУ с контрастными структурами, задач Коши для ОДУ с сингулярностями, одномерных уравнений Максвелла в слоистых диспергирующих средах

– с применением современных компьютерных технологий (п. 2 паспорта специальности 1.2.2).

Предложенные эффективные численные методы реализованы в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента: GACK-GEAD для расчетов кинетики химических реакций, SiDiaG для исследования ближайшей вещественной сингулярности в решении системы ОДУ, Continuation для численного интегрирования ОДУ с несколькими вещественными алгебраическими особыми точками, BiSpDec для решения уравнений Максвелла при наклонном падении плоской волны на набор плоскопараллельных диэлектрических пластин и исследования динамики связанных состояний (п. 3 паспорта специальности 1.2.2).

Проведены комплексные исследования научных проблем с применением новейших методов математического моделирования и вычислительного эксперимента: моделирование кинетики реакций водород-кислородного горения, расчеты спектров реальных фотонных кристаллов, формирование и динамика поверхностных волн Блоха в диэлектрическом фотонном кристалле (п. 8 паспорта специальности 1.2.2). Тем самым, в работе присутствуют оригинальные результаты одновременно из трех областей: математического моделирования, численных методов и комплексов программ.

Публикации

По теме диссертации всего опубликована 32 работы в журналах, входящих в перечень ВАК и/или индексируемых Web of Science, Scopus.

1. А.А. Белов. Численное обнаружение и исследование сингулярностей решения дифференциальных уравнений // *ДАН.* — 2016. — Т. 468, вып. 1. — С. 21–25.
2. А.А. Белов, Ж.О. Домбровская. Бикомпактная разностная схема для уравнений Максвелла в слоистых средах // *ДАН.* — 2020. — Т. 492. — С. 15–19.

3. А.А. Белов, Н.Н. Калиткин. Метод инверсной функции для задач Коши с полюсами первого порядка // *ДАН.* — 2020. — Т. 491, вып. 1. — С. 102–106.
4. А.А. Белов, Н.Н. Калиткин, П.Е. Булатов, Е.К. Жолковский. Явные методы расчета жестких задач Коши // *ДАН.* — 2019. — Т. 485, вып. 5 — С. 20–24.
5. А.А. Белов, Н.Н. Калиткин, И.П. Пошивайло. Геометрически-адаптивные сетки для жестких задач Коши // *ДАН.* — 2016. — Т. 466, вып. 3. — С. 276–281.
6. A.A. Belov, Zh.O. Dombrovskaya, A.N. Bogolyubov. A bicompact scheme and spectral decomposition method for difference solution of Maxwell's equations in layered media //*CAMWA.* — 2021. — Vol. 96C. — p. 178–187.
7. A.A. Belov, Zh.O. Dombrovskaya. The Optical Path Method for the Problem of Oblique Incidence of a Plane Electromagnetic Wave on a Plane-Parallel Scatterer // *Mathematics.* — 2023. Vol. 11. —No. 2. —p. 466.
8. A.A. Belov, N.N. Kalitkin, I.A. Kozlitin. Refinement of thermonuclear reaction rates // *Fus. Eng. Des.* — 2019. — Vol. 141. — P. 51–58.
9. А.А. Белов, Н.Н. Калиткин. Экономичные методы численного интегрирования задачи Коши для жестких систем ОДУ // *Дифф. уравнения.* — 2019. — Т. 55, вып. 7. — С. 907–918.
10. А.А. Белов, Н.Н. Калиткин. Численное интегрирование задач Коши с полюсами целого порядка // *Дифф. уравнения.* — 2022. — Т. 58, вып. 6. — С. 813–833.
11. А.А. Белов. Численная диагностика разрушения решений дифференциальных уравнений // *Ж. вычисл. матем. и матем. физ.* — 2017. — Т. 57, вып. 1. — С. 91–102.
12. А.А. Белов, Ж.О. Домбровская. Прецизионные методы решения одномерных уравнений Максвелла в слоистых средах // *Ж. вычисл. матем. и матем. физ.* — 2022. — Т. 62, вып. 1. — С. 90–104.

13. А.А. Белов, Ж.О. Домбровская. Тестирование бикомпактных схем для одномерных уравнений Максвелла в слоистых средах // *Ж. вычисл. матем. и матем. физ.* — 2022. — Т. 62, вып. 7. — С. 61–79.
14. А.А. Белов, Н.Н. Калиткин. Выбор шага по кривизне для жестких задач Коши // *Матем. моделирование.* — 2016. — Т. 28, вып. 11. — С. 97–112.
15. А.А. Белов, Н.Н. Калиткин. Особенности расчета контрастных структур в задачах Коши // *Матем. моделирование.* — 2016. — Т. 28, вып. 10. — С. 97–109.
16. А.А. Белов, Н.Н. Калиткин, Л.В. Кузьмина. Моделирование химической кинетики в газах // *Матем. моделирование.* — 2016. — Т. 28, вып. 8. — С. 46–64.
17. А.А. Белов, Н.Н. Калиткин. Проблема нелинейности при численном решении сверхжестких задач Коши // *Матем. моделирование.* — 2016. — Т. 28, вып. 4. — С. 16–32.
18. А.А. Белов, Ж.О. Домбровская. Прецизионные методы расчета нестационарных задач интегральной фотоники // *Известия РАН. Сер. физ.* — 2022. — Т. 86, вып. 2. — С. 259–265.
19. Zh.O. Dombrovskaya, A.A. Belov. Difficulties faced by Yee's scheme in photonics problems // *Journal of Physics: Conference Series.* — 2020. — Vol. 1461. — P. 012032.
20. Zh.O. Dombrovskaya, A.A. Belov, V.A. Govorukhin. Adaptive mesh for computation of electromagnetic wave propagation through high refractive index dielectric structures // *Journal of Physics: Conference Series.* — 2020. — Vol. 1461. — P. 012031.
21. A.A. Belov, N.N. Kalitkin. Error estimations for the regularized double period method // *Progress In Electromagnetics Research Symposium – Spring (PIERS), 2017*, — P. 2123–2130.

22. Belov A.A., Korpusov M.O. Numerical Blow-up Diagnostics for Differential equation solutions // *Progress In Electromagnetics Research Symposium – Spring (PIERS), 2017.* — 2017. — P. 2637–2643.
23. А.А. Белов. Пакет GACK для расчета химической кинетики с гарантированной точностью // *Препринты ИПМ им. М.В. Келдыша.* — 2015. — вып. 71.
24. А.А. Белов, Н.Н. Калиткин. Численное интегрирование задач Коши с особыми точками // *Препринты ИПМ им. М.В. Келдыша.* — 2020. — вып. 76.
25. А.А. Белов, А.С. Вергазов, Н.Н. Калиткин. Контроль точности при численном интегрировании жестких систем // *Препринты ИПМ им. М.В. Келдыша.* — 2020. — вып. 88.
26. А.А. Белов, П.Е. Булатов, Н.Н. Калиткин. Сравнительный анализ алгоритмов автоматического выбора шага для жестких задач Коши // *Препринты ИПМ им. М.В. Келдыша.* — 2019. — вып. 146.
27. А.А. Белов, А.С. Вергазов, Н.Н. Калиткин. Погрешность численного решения жестких задач Коши на геометрически-адаптивных сетках // *Препринты ИПМ им. М.В. Келдыша.* — 2019. — вып. 138.
28. А.А. Белов, О.В. Вальяников, Н.Н. Калиткин. Численное решение задач Коши с сингулярностями // *Препринты ИПМ им. М.В. Келдыша.* — 2019. — вып. 121.
29. Е.К. Жолковский, А.А. Белов, Н.Н. Калиткин. Решение жестких задач Коши явными схемами с геометрически-адаптивным выбором шага // *Препринты ИПМ им. М.В. Келдыша.* — 2018. — вып. 227.
30. П.Е. Булатов, А.А. Белов, Н.Н. Калиткин. Расчет химической кинетики явными схемами с геометрически-адаптивным выбором шага // *Препринты ИПМ им. М.В. Келдыша.* — 2018. — вып. 173.

31. А.А. Белов, Н.Н. Калиткин. Численные методы решения задач Коши с контрастными структурами // *Моделирование и анализ информационных систем.* — 2016. — Т. 23, вып. 5. — С. 528–537.

32. A.A. Belov, N.N. Kalitkin. Numerical solution of Cauchy problems with multiple poles of integer order // *Discrete & Continuous Models & Applied Computational Science.* — 2022. — Vol. 30, no. 2. — P. 105–114.

Программы решения стационарной и нестационарной задач для системы одномерных уравнений Максвелла имеют свидетельства о государственной регистрации в Реестре программ для ЭВМ за номерами № 2022663076 от 11.07.2022 и № 2022661873 от 28.06.2022.

Краткое содержание работы

Диссертация состоит из введения, 8 глав и заключения. Общий объем диссертации: страниц 383, рисунков 124, таблиц 6. Список литературы включает 531 наименований.

Введение содержит общую характеристику работы. Обоснована актуальность решаемых задач, сформулированы цели и задачи работы, указана научная новизна полученных результатов, их теоретическая и практическая значимость. Описаны методы и методология исследования, методы обоснования достоверности результатов и их апробации. Сформулированы положения, выносимые на защиту.

В главе ***Исторический очерк метода конечных разностей*** дан обзор методов конечных разностей для ОДУ, уравнений в частных производных и миметических схем. Приведено изложение метода сгущения сеток и апостериорных оценок точности в современной формулировке. Перечислены задачи, которые по-прежнему представляют трудности для существующих конечно-разностных подходов. К этой группе задач относятся ***1°*** жесткие задачи Коши с контрастными структурами, ***2°*** задачи Коши, решения которых имеют син-

гулярности на отрезке интегрирования, $\mathcal{Z}^o$ задачи для уравнений в частных производных в слоистых средах. Эти задачи имеют общую черту: их решения содержат особенности. Особенностью может быть как особая точка в узком смысле (в которой решение обращается в бесконечность), так и сильный либо слабый разрыв (в котором решение остается конечным). Контрастная структура также может рассматриваться как особенность, поскольку в пределе при увеличении жесткости она стремится к сильному разрыву.

В главе ***Жесткие задачи Коши*** предложен новый метод автоматического выбора шага – геометрически-адаптивные сетки. Он основан на использовании кривизны и наклона интегральной кривой. Показано, что зависимость шага геометрически-адаптивных сеток от кривизны является асимптотически оптимальной в смысле метрики Хаусдорфа. Разработана процедура сгущения сеток, позволяющая применить метод Ричардсона и находить апостериорную асимптотически точную оценку погрешности полученного решения (для традиционных алгоритмов выбора шага не найдено таких оценок). Поэтому предложенные методы существенно превосходят по надежности и достоверности результатов расчетов ранее известные алгоритмы.

Для расчетов по явным схемам построены экономичные методы вычисления кривизны. Для явных схем Рунге-Кутты с числом стадий до 4 приведены таблицы коэффициентов этих формул. Это позволяет успешно применять явные схемы даже к сверхжестким задачам.

Проведена апробация предложенных методов на представительных тестовых задачах. Показано, что явные схемы на геометрически-адаптивных сетках не уступают неявным схемам в надежности и точности, но кардинально превосходят их в экономичности. Проведены расчеты тех же тестовых задач по стандартным программам Гира и Дормана-Принса. Эти расчеты показали, что фактическая точность этих программ на много порядков отличается от заданной. Это иллюстрирует преимущества предложенных методов.

Проанализированы достоинства и недостатки численных методов и асимптотических методов разложения по малому параметру. Определены области их применимости и даны практические рекомендации. Приведены примеры расчетов, иллюстрирующих эти выводы.

В главе ***Кинетика химических реакций*** предложен новый специализированный численный метод для задач кинетики. Этот метод явный, и его трудоемкость очень мала. Показано, что для данного типа задач этот метод является более точным и надежным, чем другие схемы первого и второго порядка точности (специализированная схема Калиткина-Гольдина и общие схемы Эйлера и Розенброка). Метод позволяет проводить вычисления одновременно с нахождением гарантированной оценки математической погрешности и пригоден к включению в газодинамические пакеты программ.

Проведены расчеты задачи кинетики химических реакций на примере водород-кислородного горения. Показано, что явные схемы Рунге-Кутты и предложенный специализированный явный метод на геометрически-адаптивных сетках успешно справляются с этими задачами.

Проведено тестирование традиционных алгоритмов выбора шага на этой задаче, причем впервые контролировалась фактическая точность расчета. В качестве контрольного метода использовались явные схемы Рунге-Кутты на геометрически-адаптивных сетках. Показано, что традиционные программы теряют надежность, не обеспечивают заданную точность, а в ряде случаев и вовсе не позволяют провести расчет.

В главе ***Задачи Коши с сингулярностями решения*** разработан новый численный метод обнаружения и исследования ближайшей сингулярности алгебраического или логарифмического типа в решениях дифференциальных уравнений. Он позволяет надежно обнаруживать наиболее важные типы особенностей и находить их характеристики с апостериорной асимптотически точной оценкой погрешности. Этот метод применен к исследованию S-режима нелинейного горения.

Предложен новый метод решения задачи Коши для системы ОДУ с последовательностью алгебраических особых точек целого порядка. Этот метод имеет строгое обоснование. Он позволяет проводить сквозной расчет через несколько особых точек и вычислять решение с высокой точностью даже в непосредственной близости от них. Метод основан на введении обобщенной инверсной функции, для которой особая точка исходного решения является простым нулем. Преимущества метода проиллюстрированы на нетривиальных тестовых задачах.

В главе ***Разностные методы для одномерных уравнений Максвелла в слоистых средах*** для системы стационарных одномерных уравнений Максвелла построена бикомпактная разностная схема, шаблон которой включает только один шаг пространственной сетки. Границы слоев выбираются в качестве узлов сетки, поэтому схема позволяет проводить расчеты обобщенных решений с разрывом решения и его производной. Консервативность схемы обеспечивает сходимость к правильному обобщенному решению.

Для ряда важных частных случаев построено явное решение системы разностных уравнений бикомпактной схемы в конечном виде. Это решение является новым. Оно применимо для однородной среды, произвольной неравномерной сетки и произвольной конфигурации объемных и поверхностных токов.

Для нестационарных задач предложен принципиально новый метод спектрального разложения, позволяющий учитывать дисперсию материала. Закон дисперсии может быть произвольным (в том числе заданным таблично). Этот подход имеет простую физическую интерпретацию.

Для предложенных методов строго доказана сходимость для произвольных неравномерных сеток и неоднородных сред.

В главе ***Верификация бикомпактных схем для одномерных уравнений Максвелла*** построены физически содержательные и сложные для расчета тестовые примеры. С их помощью показано, что разностные методы, предло-

женные в предыдущей главе, кардинально превосходит существующие подходы по точности и надежности.

Проведены расчеты задачи о нормальном падении плоской волны на одномерный фотонный кристалл. Для учета флуктуаций толщин слоев образца предложен новый подход, названный методом виртуального эксперимента. Рассчитан спектр прохождения через реальный фотонный кристалл и проведено сравнение с ранее измеренным экспериментальным спектром. Показано, что результаты расчета хорошо согласуются с экспериментом в пределах погрешностей последнего 2-5%.

В главе ***Метод оптических путей*** рассмотрена задача о наклонном падении плоской волны на одномерный плоско-параллельный рассеиватель. Эта задача является двумерной. Предложен новый метод интегрирования уравнений Максвелла вдоль оптического луча (метод оптических путей), который позволяет рассчитывать эту задачу с помощью одномерных схем. Это существенно снижает трудоемкость решения многих актуальных задач.

Проведены расчеты тестовых задач с известным точным решением (наклонное падение плоской волны на плоскую границу раздела, интерферометр Фабри-Перо, спектры отражения и прохождения модельных фотонных кристаллов). Эти расчеты убедительно верифицируют метод оптических путей.

Проведены расчеты спектров отражения и прохождения реальных фотонных кристаллов при наклонном падении, и проведено сравнение этих расчетов с ранее опубликованными экспериментами. Показано, что предложенные подходы (бикомпактные схемы, метод оптических путей и метод виртуального эксперимента) обеспечивают хорошее согласование расчетов с экспериментом в пределах точности последнего $1 \div 7\%$.

Проведены расчеты реальной задачи о формировании поверхностной волны Блоха при наклонном падении импульса на одномерный фотонный кристалл. Проведено исследование параметров этой волны (яркость свечения и время жизни связанного состояния) в зависимости от толщин слоев фотонного

кристалла. Результаты этих расчетов можно использовать для экспериментальной реализации долгоживущих связанных состояний. Методы, предложенные в данной работе, позволили существенно повысить точность этого исследования.

В главе ***Скорости реакций*** рассмотрены некоторые задачи, которые были решены соискателем, но выходят за рамки данной работы. Предложены новые методы регрессии экспериментальных данных, измеренных со значительными погрешностями. Эти методы позволили получить статистически достоверные доверительные коридоры аппроксимирующих кривых. С помощью этих методов были найдены ***1°*** сечения и скорости 7 термоядерных реакций с участием изотопов водорода и гелия, наиболее важных в проблеме управляемого термоядерного синтеза, ***2°*** 20 химических реакций водородо-воздушного горения, существенных для задач плазмохимии (т.е. при температурах до 1-2 кК и давлениях 1 атм.), ***3°*** 15 химических реакций, описывающих автокаталитический механизм пиролиза этана при давлениях от 1 до нескольких атмосфер. Эти результаты не выносятся на защиту.

В ***Заключении*** сформулированы основные результаты работы.

1. Исторический очерк метода конечных разностей

1.1. Обыкновенные дифференциальные уравнения

1.1.1. Задачи

Начиная с XVII века появляются задачи, связанные с возникновением небесной механики и дифференциального исчисления. Они были принципиально новыми для науки и имели большую практическую значимость. Примером может служить построение геодезической сети на поверхности Земли, что было чрезвычайно важно для навигации.

Аккуратный расчет движения небесных тел потребовал составления таблиц тригонометрических функций и логарифмов.

Развитие артиллерии и стрелкового оружия и повышение дальности стрельбы предъявляло все более высокие требования к точности расчета задач баллистики. Стрельбу приходилось вести по движущимся целям (например, кораблям), причем на все бóльшие расстояния. Необходимо было учитывать такие факторы как сопротивление воздуха, движение цели, скорость и направление ветра, вращение снаряда, прецессию оси его вращения, суточное вращение Земли и т.д.

С развитием теории специальных функций потребовались конструктивные методы их расчета. Эта задача осложнялась тем, что многие специальные функции имеют сингулярности (например, полюсы), положение которых не всегда известно заранее. Из-за этого вблизи сингулярности точность резко падает. Поэтому исследованию особенностей решения дифференциальных уравнений уделялось много внимания.

1.1.2. Ранние методы

Новые задачи уже не удавалось решить аналитически, и требовались принципиально новые подходы. Одним из таких подходов стал метод конечных разностей. Суть его в том, что на очередном шаге по времени производную заменяют конечной разностью и затем решают уравнение относительно значения искомой функции в более поздний момент времени. Прообраз этого подхода просматривается в доказательстве разрешимости начальной задачи, предложенном И. Бернулли, и в работах Тейлора по степенным рядам [1]. Отчетливо метод конечных разностей был описан Эйлером [2]. Он построил явную разностную схему (схему Эйлера) для решения обыкновенного дифференциального уравнения (ОДУ). В частности, он применил ее к численному решению уравнения Риккати, которое не удавалось проинтегрировать в элементарных функциях [3].

Как известно, решение уравнения Риккати имеет подвижные полюса [4]. Эйлер обнаружил, что при приближении к этим полюсам точное и приближенное решение расходятся все заметнее до полной потери сходства после перехода через полюс. По результатам этих расчетов был сделан вывод, что метод конечных разностей вообще не пригоден для исследования подвижных особенностей решений дифференциальных уравнений. Поэтому вплоть до 2000-х годов для исследования особенностей применялись методы, основанные на разложении в ряды.

К середине XIX века точность схемы Эйлера зачастую становилась недостаточной. Расчеты велись вручную, это сильно ограничивало приемлемое количество шагов численного расчета. Поэтому возникла необходимость в методах более высокого порядка точности. В 1850-е годы Адамс предложил многошаговую схему (метод Адамса) [5]. Ее порядок точности равнялся числу шагов, используемых в схеме. При этом для асимптотически устойчивых решений погрешность росла с течением времени медленно. Это позволило применять схему Адамса, частности, для расчета траекторий небесных тел. Впоследствии

было разработано большое количество различных многошаговых методов [5]: Нюстрема, Милна, Нордсика, Лобато, Мерсона и ряд других. Каждый из них содержал несколько конкретных схем, различающихся порядком точности и другими свойствами.

Многошаговые схемы были довольно громозкими, поэтому со временем от них отказались в пользу одношаговых. В качестве примера приведем классическую одношаговую явную схему Рунге и Кутты 4-го порядка точности [6]. Она оказалась столь удачной, что до сих пор остается стандартом по умолчанию во многих распространенных пакетах программ. Например, эта схема с успехом используется в расчетах траекторных задач в рамках общей теории относительности. Отметим недавние работы А.Н. Цирулева, посвященные этим вопросам [7, 8].

Одновременно с развитием численных алгоритмов развивалась вычислительная техника. Так, во время Первой мировой войны появились первые механические, а затем и аналоговые вычислительные приборы. Это резко повысило скорость расчетов и позволило эффективно решать разностные уравнения. Одновременно математики поняли, что большинство дифференциальных уравнений, возникающих в механике и физике, не допускают аналитического решения ни в каком смысле. При этом численные методы, созданные вне концепции конечных разностей, далеко не всегда сходятся. Так, Вейерштрасс в знаменитом ныне Трактате Пуанкаре по небесной механике обратил особое внимание на критику численных методов небесной механики XIX века [9, с. 967–976].

В начале XX века выходит ряд руководств и монографий (например, работы Селиванова [10], Маркова [11], Крылова [2], Рунге [6]), в которых восстанавливаются классические результаты Бернулли и Эйлера в исчислении конечных разностей. Это закрепило метод конечных разностей как один из важнейших вычислительных методов.

По мере развития вычислительной техники происходила переоценка пригодности тех или иных численных методов. Первые компьютеры, появившиеся в

середине 1940-х годов, имели очень малую скорость и объем оперативной памяти. Поэтому тогда важнейшим требованием к методам была экономичность, то есть необходимый объем вычислений должен быть как можно меньшим. По мере возрастания мощности компьютеров экономичность отходила на второй план, а на первый план выдвигалась ***надежность*** метода. Под надежностью подразумевается совокупность следующих свойств [12]:

- правильное качественное поведение численного решения, даже если шаги сетки или другие параметры расчета выбраны не слишком удачно;
- возможность проведения расчетов большого объема без человеческого контроля;
- получение оценки погрешности одновременно с результатом.

1.1.3. Жесткость

Во второй половине XX века бурное развитие физики и техники привело к появлению новых задач, отличительной чертой которых было наличие нескольких сильно различающихся характерных временных масштабов решения. Это свойство задач принято называть ***жесткостью***.

Разномасштабность по времени обусловлена наличием в исходной физической системе процессов, протекающих с сильно различающимися скоростями. Однако нередко возникает пространственная разномасштабность, при которой решение уравнения в частных производных резко меняется при изменении пространственных координат. Примерами являются насыщение поверхностного слоя стали азотом (что приводит к упрочнению); диффузия магнитного поля в сжимающую оболочку в магнитокумулятивных генераторах сверхсильных полей; поверхностный индукционный нагрев при закалке стальных деталей; поверхностное легирование полупроводников донорами и акцепторами. Таким образом, жесткость – это свойство не только ОДУ, но и УрЧП.

Далее рассматривается понятие жесткости применительно к ОДУ. Строгое определение жесткой задачи Коши, принятое в данной работе, сформулировано в п. 1.5.1.

Типичным примером жестких задач является кинетика реакций [13]. Наименьший временной масштаб есть время протекания самой быстрой реакции, а наибольший – самой медленной. Кажется естественным отбросить медленные реакции и упростить систему. Однако если они отвечают за наработку промежуточных компонент (например, носителей цепи), то отбрасывание неправомерно. Такая ситуация типична для многих систем химических реакций.

Жесткими являются задачи с сингулярностями решения. По мере приближения к сингулярности решение резко нарастает в течение небольшого отрезка времени. Это также можно рассматривать как проявление временной разномасштабности. Примерами таких задач являются модели плазменных неустойчивостей, приводящих к пробою [14], некоторые модели нелинейного горения [15], кумуляция ударных волн на центр в мишенях термоядерного синтеза [16], ряд моделей нелинейной акустики и оптики [17, 18], механики сплошной среды [19] и т.д. Наличие сингулярности говорит о том, что исходная модель теряет применимость. Поэтому в таких задачах остро встает вопрос об обнаружении и исследовании сингулярностей.

Перечисленные задачи с сингулярностями описываются нестационарными нелинейными уравнениями в частных производных. Для их решения распространен так называемый метод прямых [12, 13]. В нем уравнение в частных производных сводят к системе ОДУ, приближая пространственный дифференциальный оператор конечно-разностным. Затем полученную систему ОДУ решают некоторой численной схемой. Эта система ОДУ является жесткой. Кроме того, для получения разумной точности приходится брать подробную сетку по пространству. Поэтому система имеет очень большой порядок (несколько сотен или даже тысяч уравнений).

Жесткими могут быть и комплексные задачи, включающие несколько процессов различной природы. Примером являются задачи моделирования верхней атмосферы, в которых учитывается многокомпонентная диффузия, теплопроводность, кинетика реакций и др.

Жесткие задачи предъявляли чрезвычайно высокие требования к точности и надежности алгоритмов. Так, Гиршфельдер и Кертисс определяли [20] жесткие задачи Коши как «уравнения, в которых определенные неявные методы, обычно дают лучший результат, обычно несравненно более хороший, чем явные методы». Вслед за ними Хайрер и Ваннер отмечают [13, с. 10], что «жесткие задачи – задачи, для которых явные схемы не работают». Поэтому был сделан вывод, что жесткие задачи требуют принципиально новых численных методов. Они бурно развиваются с 1950-х годов. К настоящему времени разработано огромное количество схем, которым посвящена обширная литература, например, [5, 13, 21–26]. Тем не менее, многие жесткие задачи до сих пор вызывают серьезные затруднения, поэтому разработка новых подходов продолжается по настоящий момент.

1.1.4. Схемы

Исторически первыми методами расчета жестких ОДУ были многошаговые неявные схемы Гиршфельдера-Кертисса. На их основе Гир разработал [27–31] программы с автоматическим выбором схемы и шага интегрирования. Они применялись для расчетов горения ракетного топлива и ряда других задач. Эти программы оказались очень удачными и широко используются до сих пор [32]. Их детали тщательно отшлифованы временем.

Подавляющее большинство современных программ основано на одношаговых многостадийных методах. В них смена шага тривиальна: каждый новый шаг не связан с предыдущим, а необходимый порядок точности достигается за счет числа промежуточных стадий. Эти схемы можно разделить на три группы: явные, явно-неявные и чисто неявные.

Явные схемы. Они не требуют вычисления матрицы Якоби правых частей. Поэтому их трудоемкость невелика. Под вычислительной трудоемкостью мы понимаем число вызовов правых частей для выполнения одного шага. Заметим, что в данной работе рассматриваются последовательные алгоритмы. Их распараллеливание выходит за рамки данной работы.

Принято считать, что явные схемы пригодны только для мягких задач. Однако это относится только к вычислениям с постоянным шагом. При использовании удачных формул выбора шага эти схемы позволяют рассчитывать немалое количество жестких задач. Хорошие результаты показывают схемы Рунге-Кутты.

Классическая схема Кутты включает 4 стадии и имеет 4-й порядок точности. Приведем современную форму записи схемы Рунге-Кутты для системы ОДУ

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}, t), \quad \mathbf{u}(0) = \mathbf{u}^0, \quad 0 < t < T. \tag{1.1}$$

где $\mathbf{u}, \mathbf{f} \in R^J$ – J-мерные вектор-функции. Чтобы опустить номер шага, обозначим через численное решение на исходном шаге через $\mathbf{u}$, а на новом шаге – через $\hat{\mathbf{u}}$. Схема с p стадиями и шагом τ имеет следующий вид:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \sum_{s=1}^{p} b_s \mathbf{w}_s, \quad \mathbf{w}_s = \mathbf{f}\left(\mathbf{u} + \tau \sum_{q=1}^{s-1} a_{sq} \mathbf{w}_q\right). \tag{1.2}$$

Бутчер предложил [33, 34] общий подход построения схем Рунге-Кутты порядка выше 4-го и впервые разработал 6-стадийную схему 5-го порядка точности. Позднее одношаговые схемы Рунге-Кутты высокого порядка точности были построены Кассити [35], Вернером [36, 37], Стоуном [38], Хаммудом [39] (в оригинальной работе имела место описка в коэффициентах, исправленная Калиткиным и учениками [40, 41]), Хашиным [42–45]. В практических вычислениях широко используется 7-стадийная схема Дормана-Принса [46–48] пятого порядка точности. На данный момент известны одношаговые схемы порядка точности вплоть до 10-го.

Явно-неявные схемы. Эти схемы имеют высокую надежность. Первое семейство таких схем с произвольным числом стадий было предложено Розенброком [49]. Запишем эту схему для случая автономной задачи

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \operatorname{Re} \mathbf{w}, \quad (E - \sigma\tau\, \mathbf{f}_{\mathbf{u}})\, \mathbf{w} = \mathbf{f}(\mathbf{u}). \tag{1.3}$$

Здесь $\mathbf{f}_{\mathbf{u}}$ – матрица Якоби правых частей. Явно-неявные схемы требуют нахождения матрицы Якоби и решения системы линейных уравнений, поэтому они в $\sim J$ раз более трудоемки, чем явные схемы.

Простейшей схемой этого класса является чисто неявная схема ROS1 с коэффициентом $\sigma = 1$. Она имеет точность $O(\tau)$.

Уникальной надежностью отличается одностадийная схема Розенброка CROS с комплекснозначным коэффициентом $\sigma = (1 + i)/2$. К сожалению, эта схема малоизвестна. Она отсутствует в монографии [13], а в учебной литературе приведена лишь в [12]. Хотя эта схема одностадийная, она имеет точность $O(\tau^2)$.

Ширков [50] и позднее Альшин и Альшина [51] построили двухстадийные схемы с комплекснозначными коэффициентами, имеющие точность $O(\tau^4)$. По надежности они несколько уступают схеме CROS, но если расчет идет без срывов, то точность оказывается существенно выше.

Различные модификации явно-неявных схем предлагались С.С. Филипповым [52, 53], Ширковым [54, 55], Новиковым [56–58].

Чисто неявные схемы. Существует ряд задач, с которыми явно-неявные схемы не справляются. Примером является задача о тепловой волне Самарского-Соболя. Она описывается квазилинейным уравнением теплопроводности, причем коэффициент теплопроводности $\varkappa$ зависит от температуры u степенным образом $\varkappa \sim u^m$. В этой задаче явно-неявные схемы дают ложную сходимость [59]. В таких случаях принципиально необходимы чисто неявные схемы.

Простейшей чисто неявной схемой является неявная (обратная) схема Эйлера

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \mathbf{f}(\hat{\mathbf{u}}). \tag{1.4}$$

В литературе описаны [13] различные реализации полностью неявных схем Рунге-Кутты (гауссовы методы, схемы Радо и Лобато, W-преобразование), большое количество диагонально-неявных методов (схемы Альта, Крузе, Нерсетта, Александера и др.), а также ряд многошаговых методов (схемы, основанные на дифференцировании назад, предиктор-корректор, метод Нюстрема и др.).

Наиболее надежными из имеющихся схем являются оптимальные обратные схемы Рунге-Кутты (Backward Optimal Runge-Kutta, BORK), предложенные Калиткиным и Пошивайло [59–61]. Это чисто неявные схемы. Они построены для числа стадий $p \leqslant 4$, при этом p-стадийная схема имеет порядок точности p и L_p-устойчивость. Схемы BORK записываются через рекурсивные функции, это позволило уменьшить их трудоемкость. Таким образом, точность и устойчивость этих схем соответствуют полностью неявным схемам Рунге-Кутты, а трудоемкость – диагонально-неявным. В качестве примера приведем схему точности $O(\tau^2)$

$$\hat{\mathbf{u}} = \mathbf{u} + 0.5\tau\mathbf{f}\left(\hat{\mathbf{u}} - 0.5\tau\mathbf{f}(\hat{\mathbf{u}})\right). \tag{1.5}$$

Аналогично записываются схемы точности $O(\tau^3)$ и $O(\tau^4)$.

Для нахождения решения на каждом шаге нужно решать систему нелинейных алгебраических уравнений. Обычно используют ньютоновский итерационный процесс. Современное состояние методов ньютоновского типа описано в монографии О. Чулуунбаатара [62]. На каждой итерации требуется вычислять матрицу Якоби и решать систему линейных уравнений вида (1.3). Поэтому трудоемкость неявных схем существенно выше, чем явно-неявных.

Специальные схемы. В [63–66] предлагались так называемые нестандартные методы, основанные на выделении в правых частях нелинейных членов и нелокальной аппроксимации.

В ряде работ (см., например, [67, 68] и цитированную литературу) предлагался экспоненциальный метод. Он основан на построении решения в виде суперпозиции экспонент, показатели которых равны элементам матрицы Якоби. Для линейных задач матрица Якоби вычисляется точно, поэтому метод име-

ет очень малую погрешность. Аналогичная ситуация имеет место для «почти» линейных задач, у которых нелинейность является малой поправкой.

1.1.5. Контроль точности

Априорные оценки. Одной из фундаментальных проблем является оценка ошибки численного метода. Лагранж вывел оценки точности многочленов Тейлора [69], Коши предложил [5, с. 166] оценки погрешности для схемы Эйлера. Рунге получил хорошо ныне известные априорные оценки методов Рунге-Кутты (см., например, [5, 70]).

Локальная оценка относится к погрешности e на одном шаге. Она имеет следующий вид. Пусть ограничены все частные производные правой части f до порядка p включительно. Тогда существует постоянная C такая, что

$$|e(h)| < C\frac{h^{p+1}}{p!}. \tag{1.6}$$

Оценка (1.6) основана на сравнении рядов Тейлора для точного решения и для разностной схемы. Постоянная C фактически есть мажоранта для p-х производных функции f.

Для глобальной погрешности имеет место следующая оценка.

Теорема 1. *Обозначим U окрестность точки $\{t, u(t) \,|\, 0 < t < T\}$, где $u(t)$ – точное решение задачи (1.1). Пусть в U Справедливы оценки погрешностей (1.6) и выполнено условие*

$$\left\| \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right\| \leqslant L. \tag{1.7}$$

Тогда имеет место оценка глобальной погрешности

$$\|E\| \leqslant h^p \frac{C}{L}(\exp(LT) - 1). \tag{1.8}$$

Теорема (1) верна как для линейных, так и для нелинейных задач.

Сделаем два важных замечания. Во-первых, фактически оценки (1.6) и (1.8) содержат высокие производные искомого решения. На практике эти производные неизвестны и – в случае жестких задач – очень велики.

Во-вторых, глобальная оценка (1.8) является мажорантной. С увеличением отрезка интегрирования T она растет экспоненциально (даже если, начиная с некоторого момента времени, точное решение перестает сколь-нибудь значимо изменяться).

Поэтому априорные оценки оказываются неконструктивны. В ряде случаев удается получить более удобные оценки. Пусть в задаче $\varepsilon du/dt = f(u)$ имеется только один малый параметр ε, отвечающий за ширину пограничных слоев. В этом случае Любих построил глобальные оценки точности для ряда схем Рунге-Кутты и Розенброка на равномерных сетках [71–75]. Эти оценки являются асимптотическими рядами по степеням ε. Они работоспособны, если ε удается определить до расчета, и если эти значения достаточно малы.

Апостериорные оценки. В стандартных программах интегрирования ОДУ, как правило, используются не априорные, а апостериорные оценки точности. В литературе предложено большое количество различных алгоритмов, которые можно разбить на две большие группы – вложенные схемы и локальное сгущение шага [5, 13]. При этом расчет производится на единственной сетке. Перед началом расчета пользователь задает желаемую точность $tolerance$ (tol). Предполагается, что погрешность этого расчета близка к tol. Опишем эти подходы.

Вложенные схемы. Этот метод наиболее употребителен. Расчет проводят по схеме порядка точности p, из промежуточных стадий которой можно составить схему порядка точности $p-1$. Эта схема называется вложенной. Вычисление каждого шага проводят по обеим схемам и сравнивают результаты. Разность этих результатов считают локальной погрешностью вложенной схемы. Если она примерно равна заданной пользователем погрешности tol (ее называют tolerance), то с этой же величиной h выполняется следующий шаг. В противном случае h увеличивают либо уменьшают по некоторому правилу. Именно так работает программа Дормана-Принса. Аналогичные подходы разрабатывали Купер [76, 77], Энгланд [78], Вернер [79], Прентис [80–82], Капс [83, 84], Кэш [85] и другие исследователи (см., например, [86–88]. Подобные алгоритмы

разрабатывались [89, 90] и для многошаговых схем, однако они громоздки и почти не применяются.

Для схем, у которых отсутствует вложенная, предлагались многошаговые оценки точности. Для этого из нескольких (2-3) шагов исходной схемы составлялась схема с порядком точности на единицу выше. Таким образом, исходная схема оказывалась «вложенной» в эту многошаговую схему. Далее применялись оценки точности и выбор шага, описанные в предыдущем абзаце. Такой подход развивали Шампин [91], Бутчер и Джакевич [92–96], Хашин [97] и другие. Эти алгоритмы очень громоздки. Кроме того, тестирование, проведенное в [94], показало, что даже на линейных задачах многошаговые оценки точности работоспособны, только если шаг интегрирования меняется не слишком сильно. Для задач высокой жесткости, напротив, типично быстрое изменение шага. В этом случае многошаговые оценки перестают соответствовать фактической погрешности расчета [94].

Локальное сгущение. В этом методе используют только одну схему и проводят вычисления с шагом h и $h/2$. Локальную погрешность определяют по разности этих двух расчетов и дальше поступают аналогично предыдущему методу. Поскольку этот метод требует вычисления двух дополнительных полушагов, его трудоемкость в среднем втрое больше предыдущего. Поэтому он менее употребителен. Этот метод реализован в программах Гира. Подобные методы для одношаговых схем Рунге-Кутты и Розенброка развивал Шампин [98, 99], Зедан [100] и ряд других исследователей [101–105].

Оригинальный алгоритм выбора шага, напоминающий локальное сгущение, был предложен в [106]. Он основан на анализе сходимости метода Ньютона при решении системы нелинейных алгебраических уравнений относительно $\hat{\mathbf{u}}$: если итерации сходятся медленно, то значение $\mathbf{u}$ (выбираемое в качестве начального приближения) далеко от $\hat{\mathbf{u}}$, и шаг следует уменьшить.

1.2. Уравнения в частных производных

Во второй четверти XX века метод конечных разностей начинают применять для решения уравнений в частных производных (УрЧП). В расчетной области вводят дискретный набор узлов, называемый сеткой. Производные, входящие в уравнение, заменяют конечными разностями с использованием некоторого набора узлов (шаблона). Это приводит к системе алгебраических уравнений относительно искомой функции, которая и называется разностной схемой. Описанный способ построения разностных схем называется методов разностной аппроксимации.

1.2.1. Задачи

В первой четверти XX века рассматривались преимущественно линейные задачи, которые были сравнительно просты. В 1940-х – 1950-х годах появились принципиально новые практические задачи, связанные с развитием атомного, термоядерного, ракетного и космического проектов, самолетостроения, ядерной энергетики и многих других. Среди них были системы нелинейных УрЧП, задачи с обобщенными решениями (например, газодинамические течения с ударными волнами), различные физические процессы (например, теплопроводности и колебаний) в слоистых средах и многие другие. Эти задачи исключительно сложны. Они способствовали наиболее интенсивному развитию метода конечных разностей для УрЧП. Кроме того, практика расчетов потребовала разработки соответствующего теоретического аппарата.

1.2.2. Устойчивость

Исследование устойчивости – по отношению к ошибкам округления, по начальным и граничным условиям и по правой части – стало ключевой проблемой теории разностных уравнений. Современное понятие устойчивости было сфор-

мулировано А.Ф. Филипповым в 1955 году [107]: для любого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, не зависящее от шагов сетки h, что если $\|\delta\phi\| < \delta$, то $\|\delta v\| < \varepsilon$. Здесь $\delta\phi$ – возмущение входных данных сеточной задачи, δv – соответствующее возмущение ее решения. Возмущение входных данных может иметь различную природу: ошибки округления в ходе вычислений, ошибка аппроксимации, физическая погрешность входных данных и т.д. Неустойчивость проявляется в нефизичном росте разностного решения с течением времени, что делает расчет практически невозможным.

Одна из первых работ по разностным уравнениям математической физики, и в том числе, устойчивости, принадлежит Куранту, Фридрихсу и Леви [108,109]. Основополагающие результаты по устойчивости разностных схем приведены в фундаментальной монографии Рябенького и Филиппова [110], некоторые из них были позднее переоткрыты Лаксом [111]. Приведем современные формулировки теорем Рябенького-Филиппова по [12].

Пусть исходная задача имеет вид $A(u(x)) = f(x)$, $x \in G$. Под x понимается совокупность всех независимых переменных. Составим для нее разностную схему $B(v(x_n)) = \phi(x_n)$, где $x_n \in \omega$ – узлы сетки, совокупность шагов которой обозначим через h. Будем считать, что схема устойчива.

Оператор A может быть дифференциальным, интегральным, интегро-дифференциальным и т.д. Оператор B является алгебраическим. Отметим, что Рябенький и Филиппов рассматривали разностную схему как некоторую абстрактную модель, не конкретизируя вид A и B, и установили важнейшие особенности поведения этой модели.

Теорема 2. *Пусть решение точной задачи существует, а разностная схема корректна и аппроксимирует точную задачу. Тогда разностное решение сходится к точному.* •

Теорема 3. *Если условия теоремы 2 выполнены, операторы A и B линейны, а порядок аппроксимации равен p, то сходимость имеет порядок не ниже p.* •

Теорема 4. *Пусть нелинейный оператор B аппроксимирует оператор A с p-м порядком и выполняется условие $\delta(\varepsilon) \leqslant \delta_0 \varepsilon^m$. Тогда имеет место сходимость с порядком $q \leqslant p/m$.* •

Теорема 5. *Пусть операторы A и B линейны, а разностная схема корректна и аппроксимирует исходную задачу с порядком p. Пусть существует функция непрерывного аргумента*

$$\overline{\psi}(x) = \lim_{h \to 0} [B(u(x_n)) - \phi(x_n)] h^{-p}.$$

Пусть также существует решение задачи

$$A\overline{z}(x) = \overline{\psi}(x), \quad x \in G,$$

и на этом решении разностный оператор B аппроксимирует дифференциальный оператор A. Тогда погрешность численного решения имеет следующую асимптотику:

$$u(x) - v(x) = h^p \overline{z}(x) + o(h^p), \quad h \to 0. \quad \bullet$$

Теорема 6. *Если для исходной задачи существует хотя бы одна корректная разностная схема, аппроксимирующая задачу на функциях $u(x) \in U$, то решение $u(x)$ исходной задачи в классе U существует и единственно. Если правая часть $\phi(x)$ непрерывна равномерно по h, то $u(x)$ непрерывно зависит от $f(x)$.* •

Теоремы 2, 4, 6 справедливы как для линейных, так и для нелинейных задач. Теоремы 3, 5 доказаны только для линейных задач. Тем не менее, на практике они успешно применяются к широкому кругу нелинейных задач (см. п. 1.4.3).

1.2.3. Схемы

По-видимому, одной из первых была схема «крест», предложенная Курантом для уравнений акустики. Схемы, подобные ей, разработаны для уравнений гидродинамики (см., например, [112–114]), уравнений Максвелла [115], а также ряда других гиперболических задач и широко применяются на практике.

Основы отечественной вычислительной математики были заложены Тихоновым и Самарским. Они построили схемы для широкого круга задач (в том числе, связанных с атомной и термоядерной проблемами). Самарский разработал [116,117] систематическую теорию разностных схем. Он выделил фундаментальные свойства разностных схем (однородность, монотонность, консервативность, экономичность и т.д.), сформулировал важнейшие методы их построения (например, интегро-интерполяционный) и обоснования (метод энергетических пространств, операторные неравенства), разработал конкретные схемы для УрЧП различного типа (эллиптического, параболического, ряда нелинейных уравнений, включая газовую динамику и др.) и методы решения разностных уравнений. Эти результаты непосредственно использовались Самарским при решении прикладных задач атомного и термоядерного проектов и ряда других. В настоящее время понятийный аппарат Самарского столь глубоко вошел во всеобщее употребление, что даже о работах Эйлера мы говорим на языке разностных схем и с использованием обозначений, предложенным Самарским.

Ряд выдающихся результатов был получен Калиткиным. Вместе с Тихоновым и Самарским он стоял у истоков отечественной школы численных методов. Калиткин разработал [12] методы решения задач, возникавших в ходе реализации термоядерной проблемы. Среди них задачи на собственные значения, уравнение переноса, параболические, гиперболические, эллиптические уравнения, ряд газодинамических задач, задачи с сингулярностями и др. Он проводил расчеты течений химически реагирующих смесей, детонационных волн, генераторов сверхсильных магнитных полей и других сложнейших прикладных задач.

Яненко построил ряд методов расчета задач механики сплошной среды, эллиптических уравнений и других проблем. Дьяконов предложил высокоэффективные разностные схемы для эллиптических уравнений, минимизирующие вычислительные затраты [118,119]. Федоренко разработал [120] сверхбыстрый метод решения многомерных краевых задач, который с успехом применялся для решения задач динамики плазмы, теории упругости, для расчёта нейтронных полей в ядерном реакторе, в задачах обтекания тел сложной формы. Годунов разработал [121] ряд методов решения уравнений газовой динамики. Среди них разностная схема, основанная на решении задачи о распаде разрыва, (широко известная «схема Годунова»), сеточно-характеристический метод, метод установления нестационарного потока, метод симметризации и др.

Для разработки надежных разностных схем для задач с обобщенными решениями принципиально важным оказалось важным учитывать законы сохранения непосредственно на стадии проектирования схем [122–124]. Примерами являются задачи газо- и гидродинамики с разрывными решениями (ударными волнами), уравнения теплоопроводности и колебаний в слоистых средах и многие другие.

Многие из перечисленных выше задач по-прежнему с трудом поддаются численному расчету. Некоторые из них удалось решить лишь в последние годы. В качестве примера приведем сильно нелинейное параболическое уравнение, описывающее тепловую волну в плотной плазме [59], нелинейные УрЧП с сингулярностями решения [125], уравнение теплопроводности и колебаний в слоистых средах [12]. Поэтому данное математическое направление продолжает бурно развиваться.

Перечислим современные методы составления разностных схем [12].

Метод разностной аппроксимации описан в начале п. 1.2. Он является наиболее распространенным.

Интегро-интерполяционный метод предложен Тихоновым и Самарским [123] и впоследствии обобщен Калиткиным (так называемые ***бикомпакт-***

ные схемы) [126, 127]. Он пригоден для уравнений с разрывными коэффициентами. Пусть сетки выбраны так, чтобы точки разрыва коэффициентов попадали в узлы. Тогда внутри каждого шага решение является гладким. Такие сетки называются специальными. Исходное уравнение интегрируют по ячейке сетки, используя квадратурные формулы (например, трапеций или средних).

Метод прямых предложен Калиткиным [128]. Этот метод применяется к нестационарным задачам (например, теплопроводности или колебаний). Он позволяет единообразно решать линейные и нелинейные задачи. Вводят сетку по пространственным переменным и аппроксимируют пространственные производные конечными разностями. В результате получают систему ОДУ огромного порядка (который равен числу точек пространственной сетки). Эту систему решают какой-либо схемой из числа описанных в п. 1.1.4.

Проекционно-сеточные методы основаны на разложении искомого решения по некоторому базису. В частности, в методе конечных элементов базисом являются кусочно-полиномиальные функции. Относительно коэффициентов разложения решают систему алгебраических уравнений, которую можно трактовать как некоторую специфическую разностную схему [12]. В методе Ритца коэффициенты разложения выбираются так, чтобы невязка была минимальна. В методе Галеркина коэффициенты выбираются так, чтобы невязка была ортогональна всем базисным функциям.

1.2.4. Асимптотические методы и гибридные схемы

Если в задаче можно выделить только один характерный масштаб ширины пограничных слоев $\varepsilon \ll 1$, то эти методы позволяют получить несложное аналитическое приближение к решению и численно его реализовать. Теория регулярных и сингулярных возмущений развивалась в работах Тихонова и его учеников (см., например, [129–138]). В настоящее время это направление активно развивается на кафедре математики физического факультета МГУ группой под руководством Нефедова и Бутузова, Быковым, Поповым и другими иссле-

дователями. Нефедов, Волков, Лукьяненко предложили [139, 140] гибридные методы, в которых решение, полученное асимптотическими методами, использовалось как априорная информация для численного расчета. В частности, это позволило построить априорно-адаптированные сетки для нестационарных задач типа адвекция-диффузия.

1.2.5. Априорные оценки точности

Метод операторных неравенств Самарского позволяет строить оценки решения по правым частям исходного уравнения и дополнительных условий [117]. Самарский применял этот подход не только для обоснования устойчивости разностной схемы, но и для получения априорных оценок точности. Пусть исходная задача $Au = f$ и разностная схема $Bv = \phi$ являются линейными. Рассматривают задачу для погрешности $z = u - v$. Эта задача имеет вид $Bz = \psi$, где ψ – невязка. Далее, пользуясь указанными оценками, мажорируют норму погрешности через некоторую норму невязки. Это позволяет построить строить оценки в сильных нормах (нередко сильнее, чем норма C) при сравнительно слабых ограничениях на гладкость правых частей и коэффициентов уравнений.

Этот подход развивали Андреев (см., например, [141–154] и другие работы этого автора), Шишкин [155, 156], Коптева [157], Стайнс [158] и ряд других исследователей.

1.3. Миметические схемы

Со временем стало ясно, что можно и нужно проектировать схемы, наследующие те или иные свойства исходной системы дифференциальных уравнений. Такие схемы получили названия «миметических» (mimetic) или «совместимых» (compatible). Описание конкретных разностных аппроксимаций этого типа и их

приложений к моделированию различных физических процессов было подытожено в [159, 160].

Из теории численного интегрирования уравнений в частных производных концепция наследования алгебраических структур возвращается в теорию обыкновенных дифференциальных уравнений в начале 1990-х годов. В частности было найдено большое семейство неявных схем Рунге-Кутты для гамильтоновых динамических систем, точно сохраняющих симплектическую структуру (см., например, [161–174] и цитированную литературу). Эти схемы сохраняют точно квадратичные интегралы, но не являются в полной мере консервативными. К тому же, эти схемы не являются явными и поэтому расчеты по ним являются весьма затратными. В современной литературе активно обсуждается эффективность таких схем и возможность построения явных схем, сохраняющих в том или ином смысле интегралы движения динамических систем [175–177].

В настоящее время вместо поиска универсальной и надежной разностной схемы для широкого круга задач рассматривается построение специализированных разностных схем, ориентированных на конкретные задачи и наследующих свойства исходной системы дифференциальных уравнений (законы сохранения, знакоопределенность решения и т.д.), важные в данной предметной области. В силу принципа универсальности математических моделей [178], методы, разработанные для моделей из одной области, могут быть применены к моделям и из другой предметной области. При этом важно, чтобы наследуемые свойства в обеих предметных областях имели ясную интерпретацию.

1.4. Расчет с контролем точности

1.4.1. Многосеточный метод

Ричардсон [179, 180] обратил внимание, что погрешность вычисления квадратур обычно имеет вид Ch^p, где h – шаг сетки, p – порядок точности разност-

ной формулы. Это позволило построить оценку точности квадратурной формулы.

Следующий шаг сделал Рунге, который применил этот подход при интегрировании ОДУ [6]. Расчет проводился двух наборе равномерных сеток, шаг которых от сетки к сетке увеличивался в два раза. Рунге сравнивал решения на двух последовательных сетках. Те десятичные знаки, которые у двух решений совпадали, считались правильными (т.н. «правило Рунге»). Рунге считал такую оценку ориентировочной [5].

Позднее Калиткин и его ученики предложили проводить расчет не с увеличением, а с уменьшением шага. При этом использовалась не ориентировочная оценка по правилу Рунге, а асимптотически точная оценка по Ричардсону, основанная на представлении погрешности как ряда по степеням h. Это позволило проводить экстраполяцию точности, причем многократно (то есть реккурентно).

Алгоритмы Ричардсона, Рунге и Калиткина весьма близки, поскольку в их основе лежит расчет не на единственной сетке, а на наборе сеток. Объединим эти алгоритмы под общим названием – многосеточный метод.

1.4.2. Задачи Коши для ОДУ

Изложим современную формулировку многосеточного метода [12, 181] применительно к интегрированию системы ОДУ (1.1) на равномерной сетке (см. также [5]).

Выберем некоторую разностную схему и выполним расчет на наборе сгущающихся сеток. Пусть первая сетка содержит N шагов величины τ, вторая сетка – $2N$ шагов величины $\tau/2$, третья – $4N$ шагов величины $\tau/4$ и т.д. При этом узлы предыдущей сетки точно совпадают с четными узлами более подробной сетки. В результате получим последовательность сеточных решений $\mathbf{u}_n^N$, $0 \leqslant n \leqslant N$, $\mathbf{u}_n^{2N}$, $0 \leqslant n \leqslant 2N$ $\mathbf{u}_n^{4N}$, $0 \leqslant n \leqslant 4N$ и т.д. Здесь индекс n соответствует номеру узла сетки по времени.

Сеточные решения зависят от шага интегрирования τ. Например, на первой из указанных сеток имеем

$$\mathbf{u}_n^N = \mathbf{u}_{ex}(t_n) + \mathbf{C}\tau^p + O(\tau^{p+1}). \tag{1.9}$$

Здесь $\mathbf{u}_{ex}(t_n)$ – точное решение, множитель $\mathbf{C}$ не зависит от τ, но зависит от конкретного момента времени t_n и выбранной схемы. Ошибкой является величина

$$\mathbf{C}\tau^p + O(\tau^{p+1}), \tag{1.10}$$

в которой при достаточно малых τ важен только первый член. Запишем выражение (1.9) для сеток с числами шагов N и $2N$, удерживая только главный член в погрешности. Получим систему равенств

$$\mathbf{u}_n^N = \mathbf{u}_{ex}(t_n) + \mathbf{C}\tau^p, \quad \mathbf{u}_{2n}^{2N} = \mathbf{u}_{ex}(t_{2n}) + \mathbf{C}\,(\tau/2)^p\,, \quad 0 \leqslant n \leqslant N. \tag{1.11}$$

Узлы с номерами n грубой сетки совпадают с четными узлами с номерами $2n$ подробной сетки. Вычитая равенства (1.11) друг из друга в совпадающих узлах, найдем поточечную оценку погрешности решения

$$\Delta_{2n}^{2N} = \mathbf{C}\left(\frac{\tau}{2}\right)^p = \frac{\mathbf{u}_n^N - \mathbf{u}_{2n}^{2N}}{2^p - 1}, \quad \Delta_n^N = 2^p \Delta_{2n}^{2N}. \tag{1.12}$$

Здесь Δ_{2n}^{2N}, Δ_n^N суть векторы размерности J, они относятся к сеткам $2N$ и N соответственно.

Вычислим для каждой компоненты ошибку $\varepsilon_j^{2N} = \Delta_{2n}^{2N}$ в норме C (или L_2) по всему профилю. Для общей характеристики погрешности усредним эту ошибку по всем компонентам

$$\varepsilon^{2N} = \frac{1}{\nu}\sqrt{\sum_{j=1}^{J} \frac{(\varepsilon_j^{2N})^2}{J}}. \tag{1.13}$$

Поскольку в реальных задачах удобнее работать не с абсолютной погрешностью, а с относительной, в формуле (1.13) введена нормировка на характерный масштаб решения ν (например, в случае кинетики реакций – суммарная начальная концентрация всех исходных реагентов).

В Российском университете дружбы народов многосеточный расчет ОДУ с апостериорным контролем точности применяется коллективом под руководством Малых. Эта группа разработала прикладной пакет программ в среде Sage [182].

Экстраполяция погрешности. Исключим оценку (1.12) из решения

$$(\tilde{\mathbf{u}}_{2n})_{2N} = (\mathbf{u}_{2n})_{2N} + (\Delta_{2n})_{2N}. \tag{1.14}$$

Сравнивая (1.10) и (1.12), получаем, что ошибка уточненного решения $(\tilde{\mathbf{u}}_{2n})_{2N}$ есть $O(\tau^{p+1})$, то есть формула (1.14) эквивалентна использованию схемы порядка точности $p+1$. Уточнение по формуле (1.14) называется экстраполяцией точности.

Подчеркнем, что приведенные выше оценки точности и экстраполяция погрешности справедливы как для линейных, так и для нелинейных задач. Приведем соответствующую теорему из [5]

Теорема 7. *Пусть некоторым методом Рунге-Кутты порядка p в результате выполнения двух шагов величины τ найдено численное значение $\hat{\mathbf{u}}$, а в результате выполнения одного большого шага длины 2τ получено значение $\tilde{\mathbf{u}}$. Тогда погрешность $\hat{\mathbf{u}}$ может быть оценена по формуле*

$$\mathbf{u}_{ex}(t_0 + 2\tau) - \hat{\mathbf{u}} = \frac{\hat{\mathbf{u}} - \tilde{\mathbf{u}}}{2^p - 1} + O(\tau^{p+2}), \tag{1.15}$$

а выражение

$$\mathbf{w} = \hat{\mathbf{u}} + \frac{\hat{\mathbf{u}} - \tilde{\mathbf{u}}}{2^p - 1} \tag{1.16}$$

аппроксимирует величину $\mathbf{u}_{ex}(t_0 + 2\tau)$ с порядком $p + 1$.

Экстраполяционное уточнение (1.14) для ОДУ, по-видимому, впервые описал [181] Калиткин в 1978 году (хотя применял задолго до этого). В западной литературе этот метод был переоткрыт в [183, 184] (см. также литературу, процитированную в этих публикациях). В первой из них дано строгое обоснование такого подхода. В второй описана его реализация и проведено сравнение на тестовых задачах с программами Гира и кодом EPISODE [185], разработанным

в Ливерморской лаборатории. Это сравнение убедительно показало преимущества метода экстраполяции погрешности.

1.4.3. Краевые и начально-краевые задачи

В случае краевых задач для ОДУ многосеточный метод реализуется точно так же, как в п. 1.4.2.

В случае УрЧП сгущение сеток нужно проводить по всем независимым переменным (и по пространству, и по времени) [186]. При этом по каждой переменной узлы предыдущей (более грубой) сетки совпадают с четными узлами следующей (более подробной сетки).

Например, пусть решение u зависит от одной пространственной переменной x и от времени t. Пусть разностная схема имеет порядок точности q по пространству и p по времени. Тогда ошибка (1.10) принимает вид $\mathbf{C}_1 h^q + \mathbf{C}_2 \tau^p + O(h^{q+1} + \tau^{p+1})$. Здесь h – шаг пространственной сетки. Отсюда видно, что сгущение сеток только по части переменных (то есть уменьшение только h либо только τ) не позволяет оценить фактическую точность. Если $p = q$, и сетки по x и t сгущаются в два раза, то нетрудно получить оценку погрешности, аналогичную (1.12)

$$\Delta^{2N}_{2n,2m} = \frac{\mathbf{u}^N_{n,m} - \mathbf{u}^{2N}_{2n,2m}}{2^p - 1}. \tag{1.17}$$

Здесь индексы m и n относятся к сеткам по x и t соответственно.

Отметим один важный нюанс. Формально обоснованием многосеточного метода для разностных схем для УрЧП является теорема 5. Она доказана только для линейных задач. Однако применение метода прямых и теоремы 7 существенно расширяет применимость многосеточного метода на нестационарные нелинейные уравнения (теплопроводности, колебаний, переноса, газодинамики и т.д.), а также на стационарные задачи (например, эллиптические), решаемые с помощью счета на установление.

1.4.4. Границы применимости

Формальная область применимости метода Ричардсона определяется теоремами 5 и 7.

Оценки (1.12) основаны на том, что старший член $\mathbf{C}\tau^p$ в формуле (1.10) является преобладающим, а последующие члены пренебрежимо малы. Для этого шаг τ должен быть достаточно мал: $\tau < \hat{\tau}$ при некотором $\hat{\tau}$. С другой стороны, при избыточно малом шаге разность в числителе (1.12) оказывается сопоставима с ошибками компьютерного округления, и оценки (1.12) перестают быть применимыми. Таким образом, либо шаг $\tau > \check{\tau}$ не должен быть слишком мал, либо разрядность чисел $K > \check{K}$ должна быть достаточно велика. Это и есть условия, при котором применимы оценки точности по Ричардсону.

Множитель $\mathbf{C}$ содержит высокие производные решения. Если имеется априорная информация, что они ограничены некоторой константой, то нетрудно получить априорные оценки на τ, при которых справедливы формулы (1.12). Однако в большинстве практических задач такая априорная информация отсутствует. Поэтому применимость ричардсоновских оценок необходимо контролировать апостериорно в ходе сгущения сеток.

Более чем полувековая практика вычислений и энциклопедический кругозор Калиткина позволили ему сформулировать практические рекомендации, при выполнении которых описанный подход работает надежно и безотказно [12, 186–188]. Они заключаются в следующем.

Вычислим такие оценки точности по каждой паре сеток. Пусть оценка ε_1 найдена по решениям на сетках с N и $2N$ шагами, оценка ε_2 – по решениям на сетках с $2N$ и $4N$ шагами и т.д. Контроль точности удобно проводить по графику зависимости ε от N, выполненному в двойном логарифмическом масштабе. Если бы в (1.10) члены $O(\tau^{p+1})$ отсутствовали, то график погрешности был прямой с наклоном, равным $-p$. Однако из-за наличия членов $O(\tau^{p+1})$ он отклоняется от указанной прямой. По мере уменьшения τ (то есть увеличения

N) это отличие уменьшается. При этом оценка (1.12) стремится к фактической погрешности, равной разности численного и точного решений. Когда погрешность становится сопоставима с ошибками округления, отличие оценки (1.12) и фактической погрешности перестает убывать.

Кривая погрешности является ломаной. Найдем наклоны ее звеньев $\tilde{p}_n = \lg(\varepsilon_n/\varepsilon_{n-1})$. Они равны фактическому порядку точности. Согласно приведенным выше рассуждениям, оценки (1.12) достоверны, если фактический порядок точности стремится к теоретическому, причем монотонно [187]. Это эквивалентно выполнению двух условий:

1. $|p - \tilde{p}_n| < |p - \tilde{p}_{n-1}|$ (откуда следует $|p - \tilde{p}_n| \to 0$),
2. $(p - \tilde{p}_n)(p - \tilde{p}_{n-1}) > 0$ (условие монотонности).

Участок теоретической сходимости можно считать начавшимся, если несколько звеньев подряд удовлетворяют этим условиям. Участок теоретической сходимости заканчивается по достижении ошибок округления, то есть при нарушении хотя бы одного из указанных условий. Эти условия легко проверяются в практических расчетах. Это позволяет писать программы, в которых сгущение сеток и контроль точности проводятся автоматически.

1.4.5. Обобщения

Метод сгущения сеток получил систематическое и всестороннее развитие в работах Калиткина и его учеников (см., например, [186]). Следуя традиции, они называли этот подход «методом Ричардсона». Был предложен ряд кардинальных улучшений этого метода.

1^o Реккурентная формулировка. Оценка точности (1.12) соответствует главному члену в разложении погрешности по степеням шага. Если исключить ее из численного решения, то главным становится следующий член $C_1\tau^{p_1}$, $p_1 \geqslant p+1$. Если ошибка (1.10) есть ряд по целым степеням шага, то $p_1 = p+1$. Существуют разностные схемы, у которых в (1.10) присутствуют только степени одинаковой четности. Для них $p_1 = p+2$.

Описанное уточнение эквивалентно расчету по разностной схеме более высокого порядка точности p_1. Проводя повторное сгущение сетки, найдем для нее оценку точности аналогично (1.12) и исключим эту оценку из решения. Результат можно рассматривать как расчет по схеме более высокого порядка точности $p_2 \geqslant p_1 + 1$ и т.д. Если метод сгущения сеток применим и ошибки округления достаточно малы, то описанная процедура кардинально повышает количественную точность расчета. Если же сетка еще не достаточно подробна для регулярной сходимости (то есть членом $O(\tau^{p+1})$ в формуле (1.10) пренебречь нельзя), то экстраполяция (в том числе реккурентная) может ухудшить точность.

Одним из первых производственных приложений этого метода были расчеты уравнений состояния сильно сжатых веществ по модели Томаса-Ферми с квантовой и обменной поправками. Они были выполнены Калиткиным в первой половине 1960-х годов [189]. Эта модель приводит к нелинейной задаче на собственные значения. Результаты этих расчетов были нужны для разработки некоторых технических конструкций, поэтому к точности предъявлялись высокие требования. Благодаря реккурентному уточнению Калиткину удалось получить точность 0.01-0.001% уже на очень скромных сетках с числом шагов $N = 256$ [190].

2^o Неизвестный порядок точности. Как правило, порядок точности исследуют, исходя из определенных предположений о гладкости решения. Если в конкретной задаче эти предположения не выполняются, то фактический порядок точности будет отличается от теоретического. Например, если гладкость решения недостаточна для применения конкретной разностной схемы, то порядок точности снижается. И наоборот, встречаются вырожденные случаи «суперсходимости», когда фактический порядок точности превосходит теоретический.

Для таких случаев Калиткин построил модификацию многосеточного метода, в которой не требуется знание априорного порядка точности. Оценка точно-

сти строится не по двум, а по трем сеткам, одновременно с ней вычисляется фактический порядок точности. Этот метод был назван процессом Эйткена [181].

***3°* Квазиравномерные сетки** были предложены Самарским в 1950-е годы и использовались для прикладных расчетов. Позднее они были опубликованы Сидоровым [191]. Квазиравномерные сетки представляют собой образ гладкого монотонного преобразования равномерной сетки и дают широкие возможности для априорного адаптирования сетки под особенности решения. Удачный выбор сетки позволяет заметно повысить точность расчета и даже проводить расчеты в неограниченной области, ставя непосредственно граничное условие на бесконечности [192]. Калиткин, Альшин, Альшина и Рогов обобщили метод Ричардсона на случай квазиравномерных сеток. По-видимому, впервые такое обобщение было реализовано в работе [190].

***4°* Неструктурированные сетки.** Многосеточный метод применим и к неструктурированным (треугольным и тетраэдральным) сеткам [127,193]. Марчук и Шайдуров описали [193] процедуру сгущения таких сеток и однократное уточнение решения. Калиткин и Корякин заметили [127], что для проведения многократной реккурентной экстраполяции необходимо использование бикомпактных схем. В [194] и [195–202] (см. также другие работы этих авторов) проводилось исследование сходимости метода конечных элементов с помощью расчетов на сгущающихся сетках. Однако контроль точности проводился на тестовых задачах при помощи сравнения с известным точным решением, но апостериорные оценки точности не вычислялись.

***5°* Другие классы задач.** Калиткин и его ученики построили обобщение метода Ричардсона на уравнения в частных производных различных типов [12, 186], некоторые виды итерационных процессов (например, поиск корня нелинейного уравнения методом секущих [187] и счет на установление для эллиптических уравнений [203]), интегральные уравнения (в том числе некорректно поставленные) [12, 204] и ряд других прикладных задач.

Таким образом, работы Калиткина и его учеников кардинально расширили область применимости многосеточного метода и сделали последний общей парадигмой разностных расчетов. Напомним, что этот подход применим для равномерных и квазиравномерных сеток с известной производящей функцией. Однако его не удается напрямую применить для адаптивных сеток, которые строятся в ходе расчета. Причина в том, что узлы последовательных сеток почти никогда не совпадают. Поэтому вопрос об оценке фактической точности таких расчетов остается открытым.

1.5. Задачи, представляющие трудности для сеточных методов

Несмотря на выдающиеся результаты, достигнутые в рамках метода конечных разностей, ряд задач по-прежнему не удается решить с помощью стандартных подходов. Приведем примеры таких задач.

1.5.1. Жесткие задачи Коши

Рассмотрим задачу Коши (1.1). Если величина $\|f\|(b-a)$ оказывается порядка 1, то задача называется мягкой. К таким задачам относится, например, задача баллистики. Такие задачи можно легко рассчитывать по классическим схемам на равномерной сетке.

Задачи, у которых $\|f\|(b-a) \gg 1$, назовем трудными. В таких задачах имеется несколько сильно различающихся характерных масштабов. Трудные задачи делятся на 3 группы.

1. Жесткие задачи, в которых интегральные кривые быстро сходятся. Примером является задача радиоактивного распада.
2. Плохо обусловленные задачи. В них интегральные кривые быстро расходятся. Примерами являются многие задачи с внутренними переходными слоями, а также ряд задач с сингулярностями решения.

3. Задачи с быстро осциллирующими решениями. Такие задачи возникают, например, в радиотехнике.

Заметим, что на практике жесткость или плохая обусловленность в чистом виде встречаются редко. Довольно часто решение может включать и жесткие, и плохо обусловленные участки.

В литературе предлагались различные определения жесткости. Широко применяются формальные определения по спектру матрицы Якоби правых частей [13, 23]: если все спектральные числа отрицательны, и среди них есть большие по модулю, то задачу относили к жестким. При этом неявно предполагается, что для линейных задач все компоненты решения затухают в соответствии с величинами спектральных чисел. Однако эти определения опровергаются примером Винограда [205, 206]. В нем для линейной неавтономной системы все собственные значения отрицательны и не зависят от аргумента t, а решение имеет экспоненциально нарастающую компоненту. Поэтому определение жесткости через спектр матрицы Якоби нельзя считать универсальным.

В ряде работ используется следующее качественное понимание жесткости: система ОДУ называется жесткой, если в ней присутствует большая разномасштабность процессов (т.е., решение имеет сильно разномасштабные участки). Этой концепции придерживались Ракитский, Устинов и Черноруцкий [26], Калиткин [12, 128], Деккер и Вервер [207], Шалашилин и Кузнецов [208] и ряд других исследователей. На основе этой концепции Ракитский предложил [26] следующее строгое определение.

Определение 1. *Система ОДУ называется жесткой на отрезке изменения независимой переменной* $[a, b]$*, принадлежащему интервалу существования ее решений, если при любых начальных условиях* $t = t_0$, $\mathbf{u}(t_0) = \mathbf{u}_0$ *и на любом отрезке* $[t_0, t_0 + \Delta t] \subset [a, b]$ *найдутся такие числа* $\tau_{ПС}$, L, N, *что*

$$\left|\frac{du_j}{dt}\right|_{t\leqslant t_0+\tau_{ПС}} \leqslant \frac{L}{N}\max_{t\in[t_0,t_0+\Delta t]}|u_j(t)|, \quad j = 1, 2, \ldots, J; \qquad t_0 + \tau_{ПС} \leqslant t \leqslant t_0 + \Delta t; \quad N \gg 1. \tag{1.18}$$

Здесь $\tau_{\text{ПС}} \ll b-a$ – ширина пограничного слоя (участка быстрого изменения решения). Производные компонент вектора $\mathbf{u}$ могут достигать величин порядка

$$L \max_{t\in[t_0,t_0+\Delta t]} |u_j(t)|. \tag{1.19}$$

Величина N показывает, во сколько раз уменьшились производные после прохождения пограничного слоя.

Определение 1 представляется наиболее адекватным. Во-первых, оно может быть единообразно применено как к скалярному ОДУ, так и к системам, в том числе полученным с помощью метода прямых. Во-вторых, это определение показывает, что жесткость задачи зависит не только от вида правых частей, но и от выбора отрезка интегрирования и начального условия. В данной работе мы будем использовать определение жесткости по Ракитскому.

Численный расчет жестких задач представляет значительные трудности. Во-первых, пограничный слой требует очень мелкого шага $\tau \sim \|f\|^{-1}$. Его необходимо выбирать по самому быстрому процессу в системе. Поэтому число шагов оказывается неприемлемо велико. Таким образом, возникает вопрос о разумной стратегии адаптивного выбора шага.

Во-вторых, недостаточно только провести расчет до заданного момента. Нужно уметь надежно оценить достигнутую точность. Как отмечалось ранее, теоретические мажорантные оценки точности при этом неэффективны: они используют значения высоких производных решения, которые априори неизвестны и для жестких задач очень велики. Поэтому особое значение приобретают асимптотически точные вычисления погрешности, проводимые одновременно с нахождением самого решения.

1.5.2. Дифференциальные уравнения с особыми точками

Многие задачи Коши для нелинейных ОДУ имеют особые точки. Последние могут быть полюсами, в том числе кратными. Кроме того, при численном расчёте серьёзную проблему представляют точки, в которых обращается в нуль пер-

вая производная или одновременно несколько последовательных производных начиная с первой. Столь же проблематичны точки, в которых сама функция конечна, но обращается в бесконечность её первая производная и, возможно, несколько последующих.

В ряде случаев заранее известно, что задача имеет только алгебраические особые точки. Примерами являются автономные задачи для одного ОДУ [70], задача многих тел [209] и ряд других. Это важная априорная информация, которая существенно упрощает решение задачи.

Задачи с сингулярностями, за редкими исключениями, не решаются в элементарных функциях. Решения ряда задач являются специальными функциями математической физики. Для практического применения они требуют тщательного исследования их свойств, но даже в этом случае решение редко удаётся довести до числа. Положение усугубляется тем, что каждый узкий класс задач требует введения своей специальной функции.

Поэтому необходима разработка численных методов, позволяющих единообразно решать задачи Коши с самыми разными типами особенностей. Численные методы должны не только находить приближённое решение, но и одновременно находить конструктивную оценку его погрешности. При этом желательно иметь не мажорантные интервальные оценки (которые могут быть сильно завышенными), а асимптотически точные.

Задачи с полюсами можно разделить на две категории.

Решение с единственной сингулярностью. Первая категория включает задачи для уравнений в частных производных и обыкновенных дифференциальных уравнений, описывающие реальные физические процессы. Примерами являются различные режимы горения термоядерных мишеней, процессы пробоя в полупроводниках, задачи нелинейной лазерной оптики и акустики, гравитационный коллапс. В некоторый момент времени в решении возникает сингулярность.

Отметим также уравнение Гросса-Питаевского, описывающее состояние Бозе-конденсата. Известно, что если псевдопотенциал межчастичного взаимодействия становится отрицательным (что соответствует отталкиванию), то это уравнение имеет сингулярные решения. Исследованию этого уравнения посвящен цикл работ Г.Л. Алфимова [210–215].

В ряде случаев (например, если сингулярность является полюсом целого порядка) возможно формально-математическое продолжение решения за полюс. Однако это не имеет практического смысла, поскольку, с физической точки зрения, в момент сингулярности исходное уравнение перестаёт описывать физическое явление. Поэтому в задачах данного типа ставится вопрос о расчете решения вплоть до ближайшей сингулярности и численном исследовании последней.

В литературе рассмотрены многие подобные задачи (см., например, [14,16]). Наиболее хорошо изучены задачи, сводящиеся к параболическому уравнению с нелинейной правой частью и в некоторых случаях – с нелинейным пространственным оператором [15]. Обзор сеточных методов для данного класса задач приведен в приложении 1.

Решение с цепочкой сингулярностей. Существует второй класс задач, которые не описывают реальный физический процесс, а являются вспомогательными математическими проблемами. Примерами являются уравнения специальных функций (гипергеометрическая функция, эллиптические функции, цилиндрические функции, гамма-функция и многие другие). Решение таких задач может иметь множественные сингулярности. Эти функции широко используются в математической физике, причем в конкретных приложениях требуются значения специальной функции не только до ближайшего полюса, но и между парой соседних полюсов. Поэтому в таких задачах требуется не только рассчитать решение вплоть до ближайшей сингулярности, но и продолжить решение за нее и рассчитать последующие сингулярности. Обзор численных методов для данного класса задач приведен в приложении 1.

Применение сеточных методов. Как отмечалось выше, численные эксперименты со схемой Эйлера убедили математиков прошлых веков в том, что метод конечных разностей не пригоден для исследования подвижных особенностей дифференциальных уравнений [2]. Поэтому теория особых точек решений дифференциальных уравнений разрабатывалась в рамках аналитической теории дифференциальных уравнений. Методами теории степенных рядов и аналитического продолжения Пенлеве доказал для ряда классов дифференциальных уравнений [4], неособых систем дифференциальных уравнений и задачи многих тел [209], что подвижные особые точки являются алгебраическими, то есть в окрестности такой особой точки $t = t_0$ решение можно разложить в ряд Пюизё

$$x = c_0 + c_1(t - t_0)^q + \dots, \tag{1.20}$$

где q – рациональное число, показатель особенности.

Вопреки устоявшемуся представлению, Марчук предложил участникам семинара в Институте вычислительной математики РАН в 2003 г. оценить положение t_0 и порядок q подвижной особой точки по методу конечных разностей. Пионерской была работа [126, 127] Калиткина, Альшиной и Корякина, основанные на одностадийной схеме Розенброка с комплексным коэффициентом (CROS) [49]. Этот метод был протестирован на дифференциальном уравнении второго порядка, обладающем свойством Пенлеве, задаче Калоджеро [216] и ряде нелинейных уравнений в частных производных [14]. Этот метод показал хорошую работоспособность, однако при применении к сложным задачам он порой оказывался недостаточно надежным и требовал высокой квалификации вычислителя. Возникла необходимость разработки более надежного и универсального метода.

Задачи со множественными сингулярностями представляют еще большую трудность. В стандартных библиотеках (например, [217]) для продолжения за полюс используются различные искусственные приемы. Прохождение цепочки полюсов представляет еще большую проблему и требует разработки специ-

альных процедур. Насколько нам известно, численные методы, единообразно решающие широкие классы таких задач, отсутствуют.

1.5.3. Слоистые среды

В слоистых средах на границах раздела свойства среды (плотность, коэффициент теплопроводности, модуль упругости, показатель преломления и т.д.) изменяются скачком. В этом случае задачи для уравнений в частных производных представляют особенную трудность. На границах раздела решение претерпевает излом либо скачок. Схемы, использующие дифференцирование через границу раздела, страдают от катастрофического падения точности. Поэтому для таких задач требуются консервативные двухточечные схемы на так называемых специальных сетках [218], у которых границы раздела являются узлами. Такие схемы называются бикомпактными. Подчеркнем, что построение бикомпактных схем основано на том, что положение особенности (разрыва решения) известно априори и не меняется в ходе расчета.

Идея бикомпактных схем восходит к классической работе Годунова [122], который предложил двухточечную схему распада разрыва. Оставалось сделать лишь небольшой шаг для формулировки концепции бикомпактности. В конце 2006 года Альшина проводила расчеты по акустическим схемам с полуцелым узлами в неограниченной области. Она обнаружила эффект отражения уходящей волны от бесконечно удаленной жесткой границы. Для решения этой проблемы Калиткин и Корякин предложили бикомпактную схему. Расчёты с её использованием показали, что эффект отражения исчезает. Эти результаты докладывались, но не были опубликованы. Затем двухточечная схема для уравнения теплопроводности в однородной среде была написана в [219], а в [126,127] была опубликована идея бикомпактности и написаны схемы для слоистых сред в одномерном и двумерном случае. Систематическое развитие концепции бикомпактных схем проводилось Роговым, Михайловской, Аристовой [220–228].

Близко к этим подходам примыкают так называемые сеточно-характеристические методы. Они были предложены Магомедовым и Холодовым (см. [229] и цитированную литературу). Этот метод успешно применялся для решения задач механики сплошных сред с границами раздела (например, распространение механических колебаний в слоистых средах, лазерное сжатие оболочек мишеней термоядерного синтеза, высокоскоростное соударение жестких ударников с деформируемыми преградами и ряд других).

Большое значение для современных физических приложений имеют задачи для уравнений Максвелла рассматриваются в слоистых средах. Для уравнений Максвелла бикомпактные схемы не разработаны. Известные конечно-разностные методы включают дифференцирование через границу раздела. Поэтому их точность оказывается недостаточна для задач практики. Подробный обзор литературы по методам решения уравнений Максвелла дан в приложении 2.

Еще одной специфической трудностью многих гиперболических задач, включая уравнения Максвелла, является то, что многие среды являются сильно диспергирующими. Это означает, что скорость распространения волны существенно зависит от ее частоты. Примерами являются ряд задач физики плазмы, распространение оптического и ИК-излучения через диэлектрические среды и др. Известные подходы не обеспечивают удовлетворительную точность для таких задач.

Важное место занимают задачи синтеза оптических наноструктур и, в частности, многослойных покрытий. Значительные результаты были получены коллективом под руководством А.В. Тихонравова, Е.Л. Гусевым, Е.Б. Ланеевым. Обзор этих результатов также приведен в приложении 2.

1.5.4. Особенности в решениях дифференциальных уравнений

Задачи, перечисленные в п. 1.5.1 – 1.5.3, имеют общую черту: их решения содержат особенности.

Традиционно под особенностью понимают обращение решения в бесконечность в некоторый конечный момент времени. Однако к особенностям можно относить также сильные и слабые разрывы.

Так, в решении жестких задач для ОДУ типично наличие контрастных структур. При увеличении жесткости контрастная структура стремится к ломаной с вертикальным звеном, т.е. возникает сильный разрыв. Его положение (за редким исключением) неизвестно до расчета. Также решение может быть непрерывно, но содержать области большой кривизны, которая увеличивается с ростом жесткости. Тогда в пределе больших жесткостей в таком решении возникает слабый разрыв. Поэтому жесткие задачи Коши с контрастными структурами следует относить к задачам с особенностями.

Как отмечалось выше, решения ряда задач для УрЧП содержат сильные или слабые разрывы. Примерами являются многие нелинейные задачи (газодинамика, квазилинейный перенос и т.д.) и задачи в слоистых средах. В первом случае положение разрыва заранее неизвестно и может меняться с течением времени. Во втором случае положение разрыва совпадает с границами раздела сред.

1.6. Результаты данной работы

В данной работе построены, обоснованы и апробированы экономичные разностные методы для следующих классов прикладных задач.

1^o Жесткие и плохо обусловленные задачи для систем ОДУ. Предложен алгоритм автоматического выбора шага по геометрическим характеристикам интегральной кривой. Разработана процедура сгущения сеток, дающая апостериорную асимптотически точную оценку погрешности расчета. Эта процедура обобщает метод Ричардсона-Калиткина на адаптивные сетки, которые строятся в ходе расчета.

Предложенные методы позволяют проводить расчеты даже сверхжестких задач по явным схемам (например, Рунге-Кутты). Особое внимание уделяется задаче кинетики реакций, для которой построена специализированная явная схема, учитывающая ряд важных особенностей задачи. Выполнены расчеты прикладной задачи кинетики водород-кислородного горения.

2° Задачи Коши с подвижными особыми точками. Эти задачи также можно отнести к жестким. Предложен алгоритм численного обнаружения и исследования ближайшей сингулярности в решении системы ОДУ. Этот метод применим также к уравнениям в частных производных, поскольку они сводятся методом прямых к системам ОДУ огромного порядка. Предложен метод расчета задач Коши для ОДУ со множественными полюсами целого порядка.

3° Задачи для системы одномерных уравнений Максвелла в слоистой диспергирующей среде. Предложена бикомпактная схема, позволяющая учитывать произвольный закон дисперсии вещества. Выполнен ряд прикладных расчетов задач полностью диэлектрической фотоники.

Перечисленные задачи важны для практических приложений и при этом крайне сложны. Их отличительной чертой является то, что применение стандартных разностных подходов сталкивается с непреодолимыми трудностями. Особенностью всех предлагаемых методов является проведение расчетов на наборе сгущающихся сеток, что обеспечивает апостериорную асимптотически точную оценку фактической точности. Результаты данной работы являются продолжением и развитием подходов, которые разрабатывались школой Калиткина на протяжении более 50 лет.

2. Жесткие задачи Коши

Традиционно рассматривают задачу Коши для системы обыкновенных дифференциальных уравнений (1.1) порядка J.

В задаче (1.1) мы не делаем каких-либо предположений о виде правых частей. Наложим только требования гладкости: будем считать, что функции f_j являются p раз непрерывно дифференцируемым. Тогда решение имеет $p+1$ непрерывную производную.

2.1. Длина дуги интегральной кривой

В задаче (1.1) можно ввести параметризацию через длину дуги интегральной кривой в пространстве переменных $\{t, u_1, \ldots, u_J\}$. Для этого формально добавим к компонентам вектора $\mathbf{u}$ нулевую компоненту, равную t. Новый вектор размерности $J+1$ обозначим через $\mathbf{U}$, $U_0 = t$, $U_j = u_j$, $1 \leqslant j \leqslant J$. Неизвестные функции $\mathbf{u}$ и аргумент t имеют разный физический смысл и разную размерность. Поэтому целесообразно ввести обезразмеривающие множители (масштабирование). В качестве такого множителя для времени выберем полное время расчета $\nu_0 = T$.

Вопрос об адекватной нормировке компонент решения намного более сложен. С одной стороны, величины компонент могут сильно отличаться друг от друга и меняться на много порядков в ходе расчета. С другой стороны, целесообразно использовать фиксированные масштабы. Мы будем нормировать компоненты решения на сумму начальных условий. Например, для задач кинетики реакций такой выбор представляется наиболее адекватным. Введем элемент длины дуги по формуле

$$dl^2 = \frac{dU_0^2}{\nu_0^2} + \sum_{j=1}^{J} \frac{dU_j^2}{\nu^2}. \tag{2.1}$$

Отсюда

$$dl = dt\sqrt{\nu_0^{-2} + \nu^{-2}(\mathbf{f}, \mathbf{f})} \equiv S dt. \tag{2.2}$$

Тогда в новом аргументе система (1.1) принимает вид

$$\frac{d\mathbf{U}}{dl} = \mathbf{F}(\mathbf{U}), \quad F_0 = \frac{1}{S}, \quad F_j = \frac{\nu f_j}{S} \quad 0 \leqslant l \leqslant L; \tag{2.3}$$

Теперь вектор правых частей является единичным, это существенно упрощает численное решение задачи.

Свойства параметризации через длину дуги интегральной кривой детально исследовали Кузнецов и Шалашилин в цикле работ, включая монографию [208]. Они ввели понятие наилучшей параметризации. Таковой называется параметризация, в которой обусловленность задачи является наилучшей. Под обусловленностью задачи Коши понимают спектральное число обусловленности соответствующей матрицы Якоби. Для случая $\nu_0 = \nu = 1$ Шалашилин и Кузнецов доказали следующее утверждение.

Теорема 8 (Шалашилин, Кузнецов)**.** *Чтобы задачу Коши для нормальной системы ОДУ (1.1) преобразовать к наилучшему аргументу, необходимо и достаточно выбрать в качестве такового длину дуги, отсчитываемую вдоль интегральной кривой этой задачи. При этом задача (1.1) преобразуется в задачу (2.3).*

Подчеркнем, что этот результат справедлив непосредственно для системы дифференциальных уравнений, а не для конкретной схемы.

Известно [208], что квадратичная погрешность, возникающая из-за возмущения элементов матрицы или правой части системы ОДУ при параметризации через длину дуги, принимает наименьшее значение. Обобщение данного результата на случай $\nu_0(t) \neq \text{const}$, $\nu(t) \neq \text{const}$ дано в [230].

2.2. Выбор шага

2.2.1. Структура решения жесткой задачи

Исторически первыми были жесткие задачи, в которых решение резко изменяется в начальный момент и дальше становится достаточно плавным. Это резкое изменение было названо пограничным слоем. Но позднее выяснилось, что такой пограничный слой может повториться через некоторое время, причем неоднократно. Такие пограничные слои были названы внутренними слоями или контрастными структурами.

На рис. 2.1, А приведен качественный вид решения жесткой задачи с контрастными структурами. По горизонтали отложен аргумент t. Обычно в структуре решения выделяют следующие характерные участки: ***1°*** области очень быстрого изменения решения, называемые пограничными слоями; ***2°*** области плавного изменения решения, которые называются регулярными.

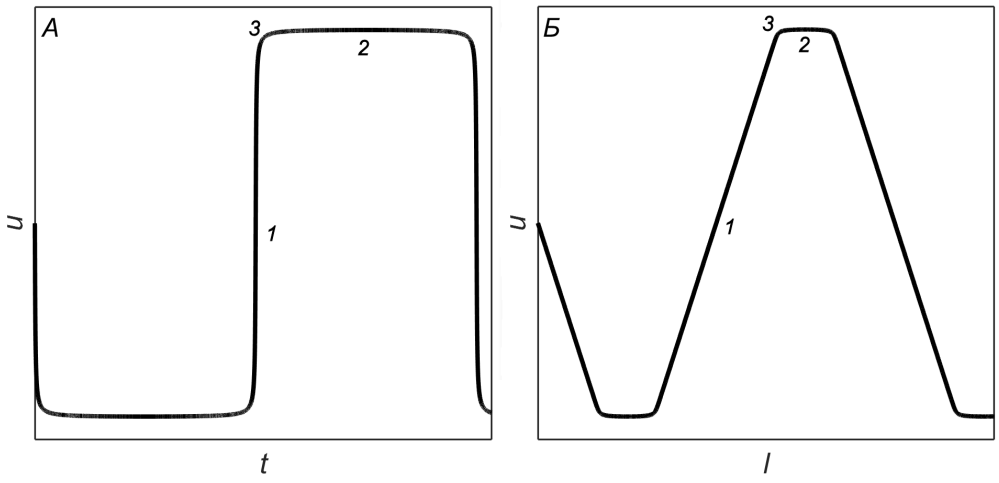

Рис. 2.1. Структура решения жесткой задачи; А – аргумент «время», Б – аргумент «длина дуги»; 1 – пограничный слой, 2 – регулярное решение, 3 – переходная зона.

Мы предлагаем выделить еще один тип характерных областей, расположенных между пограничными слоями и регулярными участками. В этих областях решение имеет большую кривизну, и мы называем их ***переходными зонами***. Кроме того, пограничный слой состоит из двух качественно разных участков

(что ранее в литературе не отмечалось). Первая половина пограничного слоя, идущая от регулярного участка, соответствует плохой обусловленности задачи. За ней следует второй участок пограничного слоя, ведущий к регулярному решению. На нем задача является жесткой в узком смысле.

Выполним переход к длине дуги интегральной кривой. Соответствующее решение приведено на рис. 2.1, Б. В этом аргументе почти вертикальные участки кривой $u(t)$ превращаются в наклонные участки кривой $u(l)$ с наклоном ± 1. Поэтому формально пограничные слои исчезают. Их, как и регулярные участки, можно рассчитывать с крупным шагом. Однако это не означает, что задача перестает быть жесткой. Переходные зоны остаются трудными для расчета.

2.2.2. Близость кривых

Традиционно близость кривых $\mathbf{u}(t)$ и $\mathbf{v}(t)$ рассматривают через норму разности $\mathbf{u}-\mathbf{v}$ [231, с. 139]. При этом используют нормы C либо L_2; в [232] использовался критерий близости кривых в норме L_1. Однако такой подход неудобен для разрывных решений, а также для жестких задач, ибо их пограничные слои очень напоминают сильные разрывы. В данной работе предложено использовать определение близости кривых, основанное на метрике Хаусдорфа.

Рассмотрим интегральную кривую $\mathbf{u}(t)$ как множество точек в евклидовом M-мерном пространстве. Расстояние между двумя кривыми определим как расстояние между соответствующими множествами в метрике Хаусдорфа [233]. Напомним это определение. Пусть U и V – два непустых компактных подмножества метрического пространства M. Пусть эти множества состоят из точек $\mathbf{u}$, $\mathbf{v}$ соответственно. Тогда расстояние от U до V есть

$$D(U,V) = \max\{\sup_{\mathbf{v}\in V}\inf_{\mathbf{u}\in U}|\mathbf{u}-\mathbf{v}|, \sup_{\mathbf{u}\in U}\inf_{\mathbf{v}\in V}|\mathbf{u}-\mathbf{v}|\} \qquad (2.4)$$

Из-за взятия операции sup определение (2.4) напоминает C. Если в этом определении заменить sup на интеграл по dl, взять $|\mathbf{u}-\mathbf{v}|$ в квадрате и извлечь из

ответа квадратный корень, то полученная метрика будет напоминать порожденную нормой L_2.

Напомним важнейшие свойства метрики Хаусдорфа [234]. Пусть $F(M)$ обозначает множество всех непустых компактных подмножеств метрического пространства M с метрикой Хаусдорфа. Тогда

- Топология пространства $F(M)$ полностью определяется топологией M.
- (Теорема выбора Бляшке) $F(M)$ компактно тогда и только тогда, когда компактно M.
- $F(M)$ полно тогда и только тогда, когда M полное.

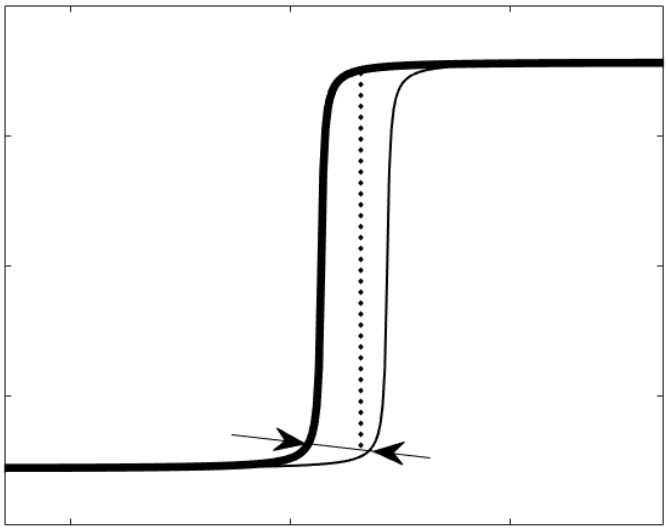

Рис. 2.2. Определение расстояния между кривыми с контрастными структурами: пунктир – традиционная разность, тонкая линия со стрелками – метрика Хаусдорфа.

Качественное сравнение предложенной метрики с традиционной разностью кривых приведено на рис. 2.2. На нем изображены две кривые с контрастными структурами, положения которых немного отличаются. Разность кривых измеряется по вертикали. Видно, что наибольшая разность примерно равна расстоянию между регулярными участками участками. Даже если положения контрастных структур отличаются на малую величину (например, $O(h)$), разность кривых не мала и составляет $O(1)$. Расстояние между этими кривыми в метрике Хаусдорфа в данном примере измеряется по почти горизонтальной линии. Видно, что оно соответствует различию в положениях контрастных структур.

Такое определение расстояния между кривыми с участками резкого изменения более адекватно, чем разность.

2.2.3. Адаптивная сетка

Выбирая шаг, мы строим на этом отрезке сетку l_n, $0 = l_0 < l_1 < \ldots < l_N = L$ из N интервалов. Из приведенных выше наглядных соображений следует, что шаг h должен быть тем меньше, чем больше кривизна $\varkappa$ интегральной кривой. Обозначим через $\varkappa_n$ кривизну и через $R_n = 1/\varkappa_n$ – радиус кривизны интегральной кривой в узлах сетки.

Наша задача – построить сетку, обеспечивающую как можно более высокую точность. Такую сетку будем называть оптимальной. Интуитивно представляется, что шаги такой сетки должны сгущаться в областях большой кривизны, но соседние шаги таких сеток не должны сильно различаться. Будем искать оптимум в классе квазиравномерных сеток.

Потребуем выполнения двух условий. Во-первых, кривизна выражается через вторые производные решения. Пусть правые части системы (2.3) имеют вторые непрерывные производные. Тогда кривизна $\varkappa(l)$ будет иметь первую непрерывную производную. Во-вторых, ограничимся классом квазиравномерных сеток l_n с дважды непрерывно дифференцируемой производящей функцией.

Построим оптимальную сетку для схемы Эйлера. Шаг по этой схеме есть движение по касательной. Сравнивая расхождение кривой и касательной на шаге h_n, получаем величину локальной ошибки на одном шаге

$$\delta_n = \frac{h_n^2}{2R_n}. \tag{2.5}$$

Сама ошибка есть вектор, перпендикулярный кривой. Таким образом, эту ошибку нужно рассматривать в смысле метрики Хаусдорфа. Тогда аналог сеточной

нормы L_2 погрешности Δ определяется выражением

$$\Delta^2 = \sum_{n=1}^{N} \delta_n^2 h_n = \frac{1}{4} \sum_{n=1}^{N} \frac{h_n^5}{R_n^2}. \tag{2.6}$$

Будем искать набор шагов h_n, минимизирующий Δ. При этом нужно учитывать, что значения R_n сами зависят от положения узлов l_n и, тем самым, от набора шагов. Удобнее приближенно перейти к непрерывному индексу n, тогда $h_n \approx dl/dn$ и $R_n = R(l)$. При сделанных выше предположениях о гладкости функций такой переход является асимптотически точным. При этом должно выполняться условие

$$\sum_{n=1}^{N} h_n \approx \int_0^N \frac{dl}{dn} = L. \tag{2.7}$$

Задача на условный экстремум $\Delta \to \min$ методом Лагранжа сводится к задаче на безусловный экстремум

$$\frac{1}{4} \int_0^N \frac{1}{R^2(l)} \left(\frac{dl}{dn}\right)^5 dn - \frac{\mu}{4} \left(\int_0^N \frac{dl}{dn} - L \right) \to \min, \tag{2.8}$$

где $\mu/4$ – множитель Лагранжа.

Вариационное уравнение Эйлера с краевыми условиями для задачи (2.8) принимает вид

$$\frac{d}{dn} \left[\frac{5}{R^2(l)} \left(\frac{dl}{dn}\right)^4 - \mu \right] + \frac{1}{2R^3(l)} \left(\frac{dl}{dn}\right)^5 \frac{dR}{dl} = 0, \quad l(0) = 0,\ l(N) = L. \tag{2.9}$$

Последнее выражение приводится к следующей форме:

$$\frac{d^2 l}{dn^2} - \frac{2}{5} \left(\frac{dl}{dn}\right)^2 \frac{d \ln R}{dl} = 0, \quad l(0) = 0,\ l(N) = L. \tag{2.10}$$

У этого уравнения нетрудно найти первый интеграл

$$h \equiv \frac{dl}{dn} = C R^{2/5}, \quad C = \text{const}. \tag{2.11}$$

Отсюда для положения узлов получаем

$$n(l) = C^{-1} \int_0^l R^{-2/5}(\tilde{l}) d\tilde{l}. \tag{2.12}$$

Константа определяется из условия $n(L) = N$. Находя эту константу, сформулируем полученный результат следующим образом.

Теорема 9. *При сделанных выше предположениях о гладкости оптимальная сетка для схемы Эйлера при $N \to \infty$ асимптотически удовлетворяет условию*

$$h_n = \frac{1}{N} \varkappa_n^{-2/5} \int_0^L \varkappa^{2/5}(l) dl, \quad 1 \leqslant n \leqslant N. \tag{2.13}$$

2.2.4. Расчетная формула

Построим расчетную формулу для шага. Обозначим через $N_{\max}$ число шагов на всей сетке с учетом переходных зон, а через $N_{\min}$ – число шагов на регулярных участках (без учета переходных зон); очевидно, $N_{\min} \ll N_{\max}$. Тогда шаг ограничивается сверху выражением

$$h \leqslant L/N_{\min}. \tag{2.14}$$

В формуле (2.11) перейдем от радиуса кривизны к кривизне и подставим явное значение константы $\text{const} = 1/N_{\max}$. В качестве расчетной формулы выберем простую интерполяцию (2.11) и (2.14)

$$h = \left[\frac{N_{\min}}{L} + \frac{N_{\max} \varkappa^{2/5}}{\int_0^L \varkappa^{2/5}(l) dl} \right]^{-1}. \tag{2.15}$$

Способ вычисления интеграла в (2.15) будет описан далее.

Очевидно, сетка, построенная указанным образом, адаптирована к решению. Будем называть ее ***геометрически-адаптивной*** (GEAD mesh – Geometrically Adaptive mesh).

Замечание. Переход от аргумента t к аргументу l кажется усложнением задачи. Однако такой переход следует делать даже для мягких задач. Поясним причину этого.

Во-первых, правые части формулы (2.3) очень просто выражаются через правые части формулы (1.1). Мягкие задачи решают явными схемами, в которых требуется вычислять только правые части (но не матрицу Якоби), поэтому для них никакого усложнения фактически не происходит.

Во-вторых, приведенный выше удачный выбор шага удалось построить только для аргумента l. Формулы выбора шага по аргументу t, используемые в методах вложенных схем или локального сгущения шага, не столь надежны, особенно в случае жестких задач.

В-третьих, в аргументе l пограничные слои перестают быть трудными участками, и легко рассчитываются крупными шагами. Измельчение шага требуется лишь в переходных зонах.

2.2.5. Вычисление кривизны

Для построения геометрически-адаптивной сетки нужно уметь вычислять вычислять кривизну многомерной кривой. Мы не встречали в литературе конструктивных формул для вычисления кривизны в многомерном пространстве. Однако есть общее определение: кривизна есть производная от единичного вектора направления касательной по длине дуги (тем самым, это вторая производная радиуса-вектора кривой по длине дуги). Реализации этого определения для явных и неявных схем различны. Опишем эти реализации.

Простейшее выражение. Вводя длину дуги в качестве аргумента, мы перешли от системы (1.1) к системе (2.3). В системе (2.3) правые части F_m суть компоненты вектора касательной к интегральной кривой. Напомним, что $\mathbf{F}$ есть вектор единичной длины. Таким образом, кривизна получается дифференцированием вектора $\mathbf{F}$ по скаляру l

$$\varkappa = \frac{d\mathbf{F}}{dl}. \tag{2.16}$$

Правые части вычисляются на каждом шаге. Напишем простейшую разностную аппроксимацию

$$\hat{\varkappa} = \varkappa(l_n) = [\mathbf{F}(u_n) - \mathbf{F}(u_{n-1})]/h_n; \quad h_n = l_n - l_{n-1}. \tag{2.17}$$

Эта аппроксимация имеет первый порядок точности, что хорошо согласуется с точностью схемы Эйлера. Поэтому такая формула пригодна для построения геометрически-адаптивных сеток.

Кривизна (2.17) вычисляется после завершения текущего шага, поэтому она может использоваться для определения величины только следующего шага. На первом шаге у нас еще нет значения кривизны. Поэтому расчет первого шага нужно повторять дважды: сначала найти величину шага по кривизне, взятой «с потолка», а по завершении шага найти кривизну и скорректировать шаг.

Явные схемы Рунге-Кутты. Выражение (2.17) по существу получено для одностадийной схемы Рунге-Кутты. В многостадийных схемах можно использовать для построения кривизны величины $\mathbf{w}_s$ из промежуточных стадий (1.2), а также величину $\hat{\mathbf{w}} = \mathbf{F}(\hat{\mathbf{u}})$. Использование $\hat{\mathbf{w}}$ не увеличивает объем расчетов: выражение для кривизны относится к следующему шагу, на котором $\hat{\mathbf{w}}$ все равно нужно вычислять. Поэтому кривизну ищем в следующем виде:

$$\hat{\varkappa} = h^{-1} \sum_{q=1}^{S+1} c_q \mathbf{w}_q, \quad \mathbf{w}_{S+1} = \mathbf{F}(\hat{\mathbf{u}}). \tag{2.18}$$

Коэффициенты c_q из (2.18) и b_q, a_{sq} из (1.2) нужно подбирать так, чтобы решение имело аппроксимацию $O(h^p)$, а аппроксимация кривизны (2.18) имела максимально возможный порядок точности. Этот анализ делается стандартным методом разложения схемы (1.2) и выражения для кривизны (2.18) по степеням h и сравнением с соответствующими разложениями для точного решения дифференциального уравнения (2.3).

c_q	b_s	a_{sq}
-1	1	-
1	-	-

Таблица 2.1. Коэффициенты схемы (1.2) и кривизны (2.18) для $p = 1$.

c_q	b_s	a_{sq}	
0	0	0	0
2	1	$1/2$	0
-2	-	-	-

Таблица 2.2. Коэффициенты схемы (1.2) и кривизны (2.18) для $p = 2$.

При этом оказывается, что для двухстадийной схемы возможно построить выражение кривизны лишь с порядком точности не выше первого. Однако при этом остается один свободный параметр, выбором которого можно уменьшить

c_q	b_s	a_{sq}		
$2/3$	$2/9$	0	0	0
-2	$1/3$	$1/2$	0	0
$-8/3$	$4/9$	0	$3/4$	-
4	-	-	-	-

Таблица 2.3. Коэффициенты схемы (1.2) и кривизны (2.18) для $p = 3$.

c_q	b_s	a_{sq}			
1	$1/6$	0	0	0	0
-2	$1/3$	$1/2$	0	0	0
-2	$1/3$	0	$1/2$	0	0
0	$1/6$	0	0	1	0
3	-	-	-	-	-

Таблица 2.4. Коэффициенты схемы (1.2) и кривизны (2.18) для $p = 4$.

коэффициент остаточного члена в кривизне. Для трех- и четырехстадийных схем возможно построить выражение кривизны лишь со вторым порядком точности. Это напоминает известные пороги Бутчера для схем Рунге-Кутты. Рекомендуемые наборы коэффициентов приведены в табл. 2.1 – 2.4.

Отметим, что нахождение кривизны по приведенным выше формулам не увеличивает трудоемкости расчетов по схемам Рунге-Кутты.

Явно-неявные схемы. По формальному определению кривизны, имеем

$$\varkappa = \frac{d\mathbf{F}}{dl} = \mathbf{F_u}\frac{d\mathbf{u}}{dl} = \mathbf{F_u}\mathbf{F}. \tag{2.19}$$

Таким образом, кривизна равна произведению матрицы Якоби от правых частей на вектор правых частей. Поскольку в явно-неявных и неявных схемах матрицу Якоби все равно приходится вычислять (это самая трудоемкая часть расчета), то попутное нахождение кривизны по формуле (2.19) не увеличивает общую трудоемкость вычислений. Поскольку матрица Якоби вычисляется ***до*** выполнения шага, то значение кривизны (2.19) можно использовать для определения величины текущего шага, а не следующего. В этом заключается качественное преимущество неявных схем перед явными.

2.2.6. Расчет с контролем точности

Расчет на единственной сетке в принципе не может дать гарантированную оценку погрешности. Единственный способ получения надежной оценки – это расчет на последовательности сгущающихся сеток и сравнение решений на этих сетках по методу Ричардсона. Этот способ дает асимптотически точное значение погрешности. Первоначально этот способ был предложен для равномерных сеток. Впоследствии было показано, что он применим на квазиравномерных сетках [12], [181], [186], а также на кусочно-равномерных и кусочно-квазиравномерных. Поэтому для гарантированной оценки погрешности нужно найти такую сетку, которая была бы геометрически-адаптивной и одновременно квазиравномерной. Для этого приходится строить процедуру расчета, состоящую из двух стадий. Опишем ее.

Построение адаптивной сетки. Это первая стадия расчета. Возьмем некоторые начальные не особенно большие значения $N_{\min}$ и $N_{\max}$ и проведем расчет по явной схеме Эйлера, используя для шага формулу (2.15). Перед началом расчета нам известно полное время T, но длина дуги L и значения интеграла (2.15) пока неизвестны. Поэтому зададим их «с потолка». Расчет на первой сетке будем вести до тех пор, пока текущее расчетное время не станет больше либо равным T. В ходе этого расчета найдем полное L вычислим значение интеграла от кривизны по любой квадратурной формуле. Затем удвоим $N_{\min}$ и $N_{\max}$, воспользуемся уже найденным значением интеграла и повторим расчет уже найденной адаптивной сетки. Она не будет сгущением первой сетки, так как ее четные узлы не будут совпадать с узлами первой сетки. Поэтому снова удвоим $N_{\min}$ и $N_{\max}$ и повторим расчет. Такое удвоение будем повторять до тех пор, пока четные узлы новой сетки не окажутся достаточно близкими к узлам предыдущей сетки.

В тестовых расчетах было опробовано несколько критериев близости сеток. Наиболее удачными оказались два следующих критерия. В первом критерии

требовалась малость величин

$$\zeta_n = (l_n - \hat{l}_{2n})(\hat{h}_{2n}^{-1} + \hat{h}_{2n+1}^{-1}). \tag{2.20}$$

Ее смысл в том, чтобы разность положений узлов должна быть мала по сравнению с соседними шагами. Здесь l относится к более грубой сетке, $\hat{l}$ – к более подробной сетке.

Во втором критерии сравнивались только отношения соответствующих шагов на двух сетках

$$\zeta_n = \sqrt{\xi_n} - 1/\sqrt{\xi_n}, \quad \xi_n = \frac{\hat{h}_{2n} + \hat{h}_{2n+1}}{h_n}. \tag{2.21}$$

В обоих случаях требовалась малость среднеквадратичной нормы этих величин

$$\|\zeta\| = \sqrt{\frac{1}{N}\sum_{n=0}^{N} \zeta_n}. \tag{2.22}$$

Данный этап сгущения проводится до тех пор, пока величина ζ не станет меньше либо равна заданному ζ_0.

Схема Эйлера используется по следующим причинам. Во-первых, она наименее трудоемка среди всех известных схем. Во-вторых, среди явных схем она наиболее надежна. Ее низкая точность несущественна, поскольку результат расчета нужен только для построения геометрически-адаптивной сетки.

Квазиравномерное сгущение сетки. Напомним, что априорные мажорантные оценки погрешности через производные решения неконструктивны и обычно даже так называемые неулучшаемые оценки сильно превышают фактические погрешности. Асимптотически точную величину погрешности дает только метод Ричардсона. Однако для применения этого метода к дифференциальным уравнениям необходимо сгущать сетки именно вдвое, причем так, чтобы все узлы предыдущей сетки совпадали с четными узлами новой сетки (тогда возможно поточечное сравнение сравнение решений на соседних сетках и вычисление сеточных норм погрешности). Сами сетки должны быть равномерными или квазиравномерными.

Не стадии построения адаптивной сетки узлы соседних сеток не совпадают. Однако если выполнен критерий совпадения сеток, то построенная сетка хорошо соответствует кривизне точного решения. Тем самым, эта сетка получается из равномерной некоторым гладким преобразованием, то есть является квазиравномерной сеткой. Поэтому ее можно взять за основу для квазиравномерного сгущения и применения метода Ричардсона.

Для квазиравномерного сгущения между каждой парой узлов старой сетки нужно поставить узел новой сетки, соответствующий той же производящей функции. Приведем формулы для нахождения такого узла. Пусть шаг h_n является внутренним интервалом исходной сетки. Тогда он делится на два интервала с шагами

$$\hat{h}_{2n-1} = h_n \frac{\sqrt[4]{h_{n-1}}}{\sqrt[4]{h_{n-1}} + \sqrt[4]{h_{n+1}}}, \quad \hat{h}_{2n} = h_n \frac{\sqrt[4]{h_{n+1}}}{\sqrt[4]{h_{n-1}} + \sqrt[4]{h_{n+1}}}. \tag{2.23}$$

Если интервал примыкает к левой границе, то его шаг h_1 делится на два новых шага по правилу

$$\hat{h}_1 = h_1 \frac{\sqrt{h_1}}{\sqrt{h_1} + \sqrt{h_2}}, \quad \hat{h}_2 = h_1 \frac{\sqrt{h_2}}{\sqrt{h_1} + \sqrt{h_2}}. \tag{2.24}$$

Для интервала h_N, примыкающего к правой границе, получаем деление

$$\hat{h}_{2N-1} = h_N \frac{\sqrt{h_{N-1}}}{\sqrt{h_{N-1}} + \sqrt{h_N}}, \quad \hat{h}_{2N} = h_N \frac{\sqrt{h_N}}{\sqrt{h_{N-1}} + \sqrt{h_N}}. \tag{2.25}$$

Таким образом, на втором этапе соответственные узлы двух последовательных сеток точно совпадают, а разность двух соседних шагов

$$\hat{h}_{2n+1} - \hat{h}_{2n} = \hat{h}_{2n+1} \left(1 - \sqrt[4]{h_{n-1}/h_{n+1}}\right) \tag{2.26}$$

есть величина более высокого порядка малости, чем шаг $\hat{h}$. Поэтому такая сетка является квазиравномерной.

Поэтому можно применить метод Ричардсона однократно к каждой паре сеток. Это позволяет проводить вычисления с автоматическим выбором шага и гарантированной оценкой погрешности, чего не было во всех ранее существовавших алгоритмах.

Проиллюстрируем описанную процедуру с помощью алгоритма 1.

Algorithm 1 Расчет на геометрически-адаптивных сетках с контролем точности

Задать $N_{\min}$, $N_{\max}$, L, $\int \varkappa dl$
while $\|\zeta\| > \zeta_0$ **do**
    while $t < T$ **do**
        Вычислить h по формуле (2.15)
        Вычислить $\hat{\mathbf{u}}, t$ по формуле (1.2)
        Вычислить $\varkappa$ по формуле (2.18)
        Вычислить ζ по формулам (2.20) либо (2.21)
    end while
    Вычислить $\|\zeta\|$ п формуле (2.22)
    $N_{\min} \leftarrow 2N_{\min}$, $N_{\max} \leftarrow 2N_{\max}$
end while
while $\Delta > tol$ **do**
    Сгустить сетку по формулам (2.23), (2.24), (2.25)
    Вычислить решение на полученной сетке, на каждом шаге пользуясь формулами (1.2)
    Вычислить поточечную оценку по формуле (1.12)
    Вычислить норму Δ оценки погрешности по формуле (1.13)
end while

Экстраполяция точности. Пользуясь найденной оценкой точности, можно провести однократную экстраполяцию погрешности, т.е. прибавить оценку точности к полученному результату. Это повышает порядок точности на 1.

Однако пользоваться реккурентным вариантом метода Ричардсона, многократно повышающим порядок точности, в данном случае затруднительно. Это связано с тем, что нечетные узлы подробной сетки $\hat{l}$, построенные по правилам (2.23) – (2.25), не точно совпадают с нечетными узлами сетки, построенной по предельной производящей функции. Эта ошибка складывается из двух частей.

Во-первых, шаги h_n грубой сетки известны не точно, а строятся по приближенной производящей функции. Чтобы можно было применять рекуррентное уточнение по Ричардсону, погрешность производящей функции должна быть меньше, чем та погрешность, которую мы рассчитываем получить в результате уточнения. Это возможно только на чрезвычайно подробных сетках, поэтому на практике не реализуется.

Во-вторых, правила (2.23) – (2.25) вносят некоторую погрешность даже при применении к точной производящей функции. Эту часть погрешности можно оценить, пользуясь определением шага квазиравномерной сетки через производную x' производящей функции в полуцелом узле

$$h_n = \frac{1}{N} x' \left(\frac{n - 0.5}{N} \right). \tag{2.27}$$

Запишем выражения для шагов h_{n-1} и h_{n+1}, аналогичные (2.27), подставим их в выражение для $\hat{h}_{2n}$ из (2.23) и разложим последнее в ряд по $1/N$ в окрестности точки $(n - 0.5)/N$. Получим

$$\hat{h}_{2n} \approx \frac{x'}{2N} \left[1 + \frac{1}{4N} \frac{x''}{x'} + O\left(\frac{1}{N^3} \right) \right], \tag{2.28}$$

где все производные берутся в точке $(n - 0.5)/N$. С другой стороны, точное выражение для этого шага имеет вид

$$\hat{h}_{2n} = \frac{1}{2N} x' \left(\frac{n - 0.5}{N} + \frac{1}{4N} \right) \approx \frac{x'}{2N} \left[1 + \frac{1}{4N} \frac{x''}{x'} + \frac{1}{32N^2} \frac{x'''}{x'} + O\left(\frac{1}{N^3} \right) \right]. \tag{2.29}$$

Сравнивая (2.28) и (2.29), найдем ошибку «ручного» дробления шага

$$\Delta\hat{h}_{2n} \approx \frac{1}{64N^3}x''' + O\left(\frac{1}{N^4}\right). \tag{2.30}$$

Таким образом, формулы сгущения (2.23) – (2.25) имеют третий порядок точности. Однако их можно применять к одношаговым схемам решения задачи Коши любого порядка точности (в том числе, выше третьего). Это объясняется тем, что в одношаговых схемах нет дифференцирования сквозь узел. Но для многошаговых схем (например, схем Адамса) такое сгущение сеток может ограничить предельный порядок точности.

2.3. Апробация методов

2.3.1. Тест

Рассмотрим тестовую задачу

$$\frac{du}{dt} = -\lambda(t)\frac{\left(u^2 - a^2\right)^2}{\left(u^2 + a^2\right)}, \quad u(0) = u^0, \quad |u^0| < a. \tag{2.31}$$

Эта задача при соответствующем выборе u^0 имеет 2 стационарных решения $u = \pm a$. Эти решения являются корнями вырожденного уравнения кратности 2. Они устойчивы, если $|u^0| < a$, и неустойчивы при $|u^0| > a$.

Задача (2.31) решается в квадратурах, что делает ее очень удобным тестом для численных расчетов:

$$u(t) = -\frac{2\,\Lambda(t)\,a^2}{1 + \sqrt{1 + 4a^2\Lambda(t)^2}}, \quad \Lambda(t) = \frac{u^0}{(u^0)^2 - a^2} + \int_0^t \lambda(\tau)d\tau. \tag{2.32}$$

Нетрудно выбрать такое $\lambda(t)$, чтобы квадратура точно вычислялась, а само решение передавало те или иные особенности. Например, возьмем $\lambda(t) = \lambda_0 \cos t$. Тогда

$$\Lambda(t) = \frac{u^0}{(u^0)^2 - a^2} + \lambda_0 \sin t. \tag{2.33}$$

На рис. 2.3 показано решение этой задачи при различных λ_0. Видно, что при увеличении λ_0 решение стремится к ступенчатому. Это наглядная интерпретация контрастных структур в случае высокой жесткости.

Для практической помощи вычислителю дадим некоторую качественную классификацию, основанную на практике численных расчетов и не претендующую на математическую строгость.

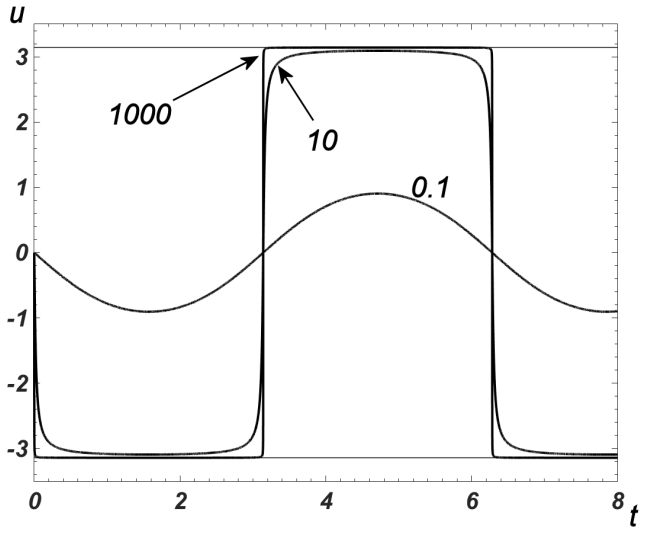

Рис. 2.3. Тест (2.31); жирные кривые – точное решение (2.32) – (2.33) при разных значениях λ_0 (указаны около кривых); тонкие прямые – стационары.

Мягкими назовем задачи, решения которых на графике хорошо отличимы от стационаров (случай $\lambda_0 = 10^{-1}$ на рис. 2.3).

Жесткими назовем задачи, решения которых с графической точностью ложатся на стационары, но визуально ширина пограничного слоя ненулевая (случай $\lambda_0 = 10^{1}$ на рис. 2.3).

Сверхжесткими назовем задачи, в которых ширина пограничного слоя визуально пренебрежимо мала, так что решение имеет ступенчатый характер; при этом регулярные участки визуально сливаются с корнями вырожденного уравнения (случай $\lambda_0 = 10^{3}$ на рис. 2.3).

Ультражесткими назовем задачи, у которых регулярные участки решения в пределах ошибок компьютерного округления неотличимы от корней

вырожденного уравнения (для данного примера это будет при $\lambda_0 = 10^5$ и более, если вычисления 64-битовые).

Категорию жесткости надо учитывать, выбирая метод решения задачи. Но по виду исходного уравнения априори определить категорию жесткости удается не всегда. Зачастую это приходится делать методом проб и ошибок.

2.3.2. Критерий качества сетки

На рис. 2.4 показана зависимость среднеквадратичной нормы величины (2.21) от числа узлов. Каждая линия соответствует определенной жесткости задачи λ_0. На задачах малой жесткости ($\lambda_0 = 1$) зависимость критерия качества от N полностью соответствует теоретическим представлениям. Расчетная кривая близка к прямой с наклоном -1. Первое звено удовлетворяет требованию монотонности, но его наклон несколько отличается от теоретического. Это является зарождением нерегулярного участка кривой.

При еще большей жесткости $\lambda = 10^3$ нерегулярный участок захватывает уже 4 первых звена. При этом первое звено кривой является нарастающим, то есть на нерегулярном участке нарушается монотонность критерия качества. При дальнейшем увеличении жесткости увеличивается как длина нерегулярного участка, так и отрезок нарушения монотонности. При этом на начальном участке возникает пилообразность.

Поясним причину возникновения возрастающего нерегулярного участка. В пределах расчетного интервала находится 3 пограничных слоя. Для мягкой задачи сетка сразу же хорошо строится во всех этих трех слоях. Однако если жесткость велика, то картина выглядит иначе.

Погрешность численного решения увеличивается к концу промежутка интегрирования. Поэтому во втором пограничном слое погрешность гораздо больше, чем в первом (начальном). Накопление погрешности в решении приводит к накоплению ошибки при вычислении шага. Поэтому сетка хорошо строится в начальном пограничном слое, ощутимо хуже – во втором слое, а в третий слой

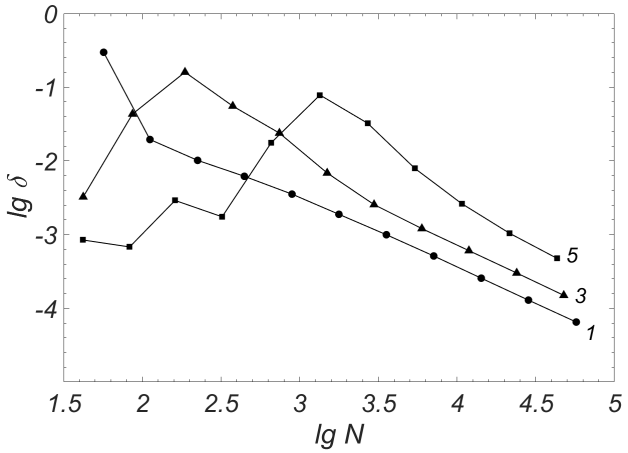

Рис. 2.4. Зависимость критерия качества (2.21) от числа интервалов сетки. Около каждой кривой указана величина жесткости $\lg \lambda_0$.

узлы сетки могут и вовсе не попадать. Поэтому на грубых сетках разностная погрешность в третьем слое вообще не вычисляется.

По мере сгущения сетки узлы начинают попадать в третий слой и в них возникает значительная погрешность. Поэтому при первых сгущениях погрешность увеличивается с увеличением числа интервалов сетки. Эти сетки образуют нерегулярный участок. При дальнейшем сгущении, когда в каждом пограничном слое становится достаточно много узлов, погрешность начинает уменьшаться в соответствии с теорией.

Для всех жесткостей переход на регулярный участок происходит при одновременном выполнении двух условий: 1) кривая монотонно убывающая (чтобы исключить начальную нерегулярность) и 2) величина критерия качества принимает значение $\delta \leqslant 10^{-2}$ для любой жесткости. Это означает, что соответственные шаги соседних сеток в среднем отличаются на $\sim 1\%$, однако максимальное отличие шагов друг от друга может быть существенно больше.

Для других критериев качества поведение кривых имеет сходный вид, однако количественное значение перехода на регулярный участок будет своим для каждого критерия.

2.3.3. Сходимость

На рис. 2.5 показаны профили решения на сгущающихся сетках для теста (2.31) с $\lambda_0 = 10^{-1}$. Для наглядности профили построены в координатах $u(t)$, хотя непосредственно рассчитывались $u(l)$ и $t(l)$. Хорошо видно, что при сгущении сеток профили стремятся к точному решению.

На рис. 2.5 показан расчет ультражесткой задачи с $\lambda_0 = 10^7$. После прохождения пограничного слоя и переходной зоны кривая выходит на стационарное решение с точностью до ошибок округления. Но уже на ближайших шагах ошибки округления выводят кривую в область неустойчивости, после чего она быстро уходит от точного решения. Видно, что численные методы непригодны для задач столь большой жесткости. В этом случае надо пользоваться асимптотическими разложениями по малому параметру.

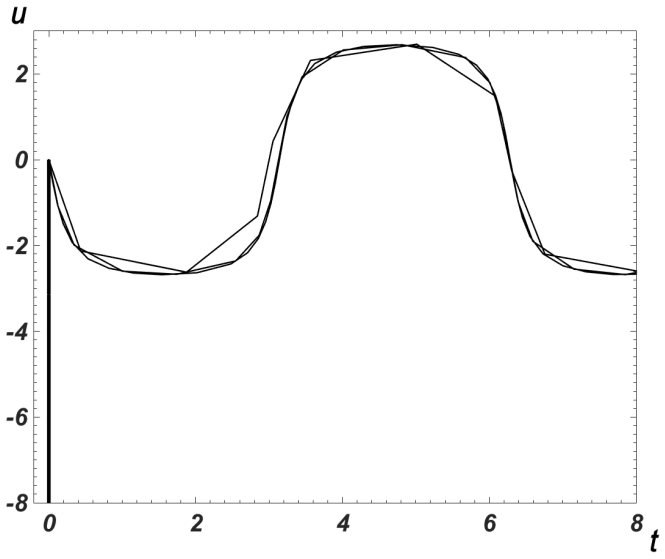

Рис. 2.5. Тонкие линии – решение теста (2.31) на сгущающихся сетках, $\lambda_0 = 10^{-1}$. Жирная линия – численное решение ультражесткой задачи с $\lambda_0 = 10^7$.

При аргументе l решение теста (2.31) не выражается в элементарных функциях, поэтому погрешность вычислялась следующим образом. Для каждой сетки l_n вычисление по разностной схеме давало приближенные значения u_n и t_n. Решение u_n сравнивалось с точным решением (2.32), вычисленным в моменты t_n. Их разности давали значения погрешности в узлах l_n. Для пересчета к аргу-

менту t использовалась формула приведенной погрешности (3.13). На график выводилась приведенная погрешность в норме L_2 в зависимости от числа узлов N в двойном логарифмическом масштабе.

На рис. 2.6 показаны погрешности по схеме ERK4 при разных λ_0. Темными маркерами на рис. 2.6 изображены погрешности, полученные путем прямого сравнения сеточного решения с точным. При $\lambda_0 = 10$ погрешность уже на грубых сетках убывает в соответствии с теоретическим порядком точности $p = 4$. При $\lambda_0 = 10^3$ кривая начинается с горизонтального участка. Это объясняется тем, что на грубых сетках сеточное решение хорошо описывает первый и второй пограничный слой, но вместо третьего пограничного слоя "срывается", давая абсурдные результаты. При увеличении λ_0 начальный горизонтальный участок удлинняется, захватывая все большее число сеток, но затем кривые выходят на участок теоретической сходимости.

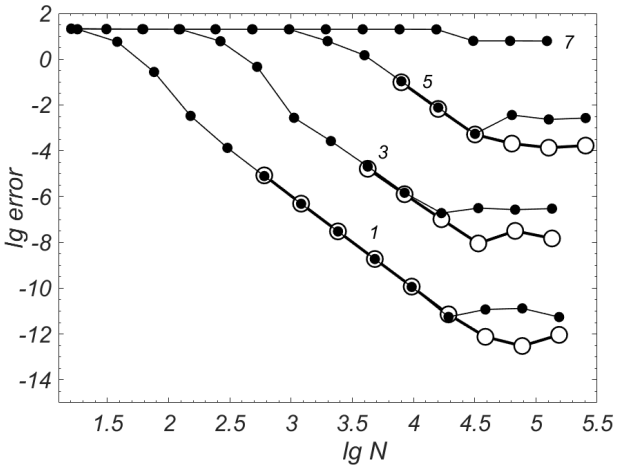

Рис. 2.6. Сходимость в тесте (2.31), у линий указаны $\lg \lambda_0$. • – погрешности, вычисленные сравнением с точным решением, ∘ – оценки по методу Ричардсона.

На достаточно подробных сетках каждая линия выходит на горизонтальный фон ошибок округления. Этот фон весьма низок при $\lambda_0 = 10$, а с увеличением λ_0 довольно быстро возрастает. Фон связан с тем, что чем больше λ_0, тем ближе точное решение подходит к стационарам. Разность между точным решением и стационарным фактически является начальными данными для следующего пограничного слоя. Если в этой разности (с учетом конечной разрядности ком-

пьютера) осталось мало достоверных знаков, это ограничивает точность дальнейшего расчета.

Если фон ошибок округления оказался неприемлемо большим, это означает, что данным методом при данной разрядности чисел мы не можем решить задачу. В этом случае нужно либо провести вычисления с повышенной разрядностью чисел, либо использовать другие подходы (например, разложение по малому параметру).

На каждой паре соседних сеток (на втором этапе сгущения) проводилась оценка погрешности по Ричардсону, не использующая сравнение с точным решением. Эти оценки показаны светлыми кружками. Видно, что на прямолинейном участке с теоретическим наклоном Они отлично совпадают с непосредственным вычислением погрешности. Это относится даже к сверхжестким задачам. Это показывает, что и для задач с неизвестным точным решением можно уверенно пользоваться ричардсоновскими оценками, если кривая погрешности содержит указанный прямолинейный участок.

На данном рисунке показан также расчет для ультражесткой задачи с $\lambda_0 = 10^7$. Видно, что прямолинейный участок погрешности с теоретическим наклоном отсутствует, поэтому для таких задач ричардсоновскую оценку погрешности дать невозможно.

2.3.4. Равномерные сетки

Влияние сгущения сетки по кривизне проиллюстрировано на рис. 2.7. На нем показана зависимость погрешности относительно точного решения от фактического числа узлов N в двойном логарифмическом масштабе. В качестве теста выбрана задача (2.31) с большим $\lambda_0 = 10^5$. Расчеты велись по явной схеме ERK4 и явно-неявной схеме Розенброка CROS.

Видно, что для схемы ERK4 на регулярном участке сходимости геометрически-адаптивная сетка дает в $\sim 300 - 1000$ раз лучшую точность, чем равномерная сетка $h_s = \text{const}$. Это обусловлено тем, что участок нерегулярной

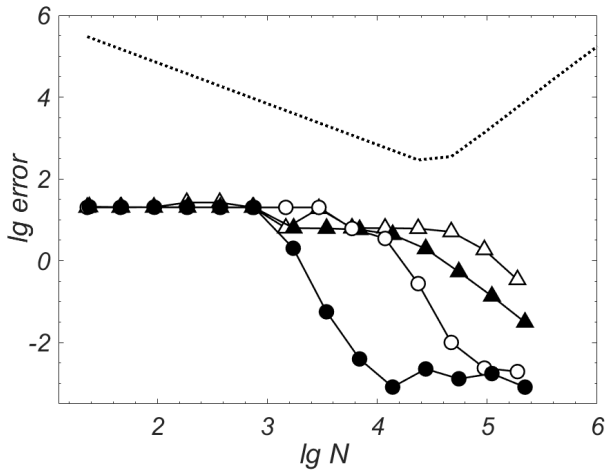

Рис. 2.7. Погрешности относительно точного решения при $\lambda = 10^5$. Схемы: $\circ$ – ERK4, $\triangle$ – CROS. Сплошные линии – расчет в аргументе l: темные маркеры – GEAD-сетка, светлые маркеры – равномерная сетка $h = \text{const}$. Пунктирная линия – сетка 2.34 в аргументе t.

сходимости оказывается более коротким, и раньше начинается быстрое убывание погрешности. На достаточно подробных сетках каждая схема выходит на фон ошибок округления, который при данном λ_0 составляет $\sim 10^{-3}$. Для схемы CROS сгущение сетки в переходной зоне также дает ощутимый выигрыш по точности – более 10 раз по сравнению с равномерной сеткой в длине дуги.

В литературе описана программа DUMKA, разработанная Лебедевым (см. [235] и процитированные там работы этого автора). Она основана явных схемах с адаптивным выбора шага и специфической апостериорной фильтрацией профилей численного решения. Эта схема применялась в НИЦ «Курчатовский институт» для расчетов задач, приводящих к огромным системам жестких ОДУ, содержащих до 10^6 уравнений [235]. Шалашилиным и Кузнецовым [208] было проведено сравнение расчетов с равномерным шагом по длине дуги и по программе DUMKA. Они показали, что переход к длине дуги обеспечивает существенно лучшую точность, чем автоматика шага в программе DUMKA. Из рис. 2.7 видно, что применение геометрически-адаптивных сеток позволяет кардинально улучшить точность по сравнению с равномерной сеткой в длине дуги. Таким образом, схема ERK4 на GEAD-сетке существенно превосходит программу DUMKA по точности.

Формулу для шага из теоремы 2.13 можно трактовать следующим образом. Переход к длине дуги эквивалентен измельчению шага по времени в пограничных слоях, а множитель $\varkappa_n^{-2/5}$ есть дополнительное сгущение шага в переходных зонах. Был проведен расчет, в котором сетка сгущалась только в переходных зонах, но не в пограничном слое. Это эквивалентно расчету в аргументе время с шагом

$$\tau_n = \tau_0 \varkappa_n^{-2/5} \tag{2.34}$$

где $\tau_0 = \text{const}$ – шаг на прямолинейном участке решения. Здесь также было выбрано $\lambda_0 = 10^5$.

Расчет по явной схеме ERK4 разваливался из-за переполнения. Схема CROS позволила завершить расчет. Полученные погрешности приведены на рис. 2.7. Видно, что измельчение шага только в переходной зоне не дает выигрыша по точности по сравнению с равномерной сеткой в аргументе t. Это показывает, что геометрически-адаптивные сетки необходимо применять только в аргументе l.

2.3.5. Вычисление кривизны

Рассмотрим три способа вычисления кривизны, которая используется для расчета шага на первом этапе: 1) по разностной формуле (2.17) первого порядка точности, 2) по разностной формуле (2.18) второго порядка точности, 3) непосредственно по кривизне точного решения по следующей формуле:

$$\begin{gathered} \varkappa = \frac{f(u,t) f_u(u,t) + f_t(u,t)}{(1 + f(u,t)^2)^{3/2}}, \\ f_u = -2\lambda \cos(\omega t) \frac{u(u^2 - a^2)(u^2 + 3a^2)}{(u^2 + a^2)^2}, \quad f_t = \lambda \omega \sin(\omega t) \frac{(u^2 - a^2)^2}{u^2 + a^2}. \end{gathered} \tag{2.35}$$

Графики нормы L_2 погрешности второго этапа для $\lambda = 10^3$ в этих случаях представлены на рис. 2.8.

Из рис. 2.8 видно, что все три способа дают примерно одинаковые результаты. Вероятная причина состоит в том, что кривизна является вспомогательной величиной для вычисления шага сетки, так что ее погрешность не сильно влия-

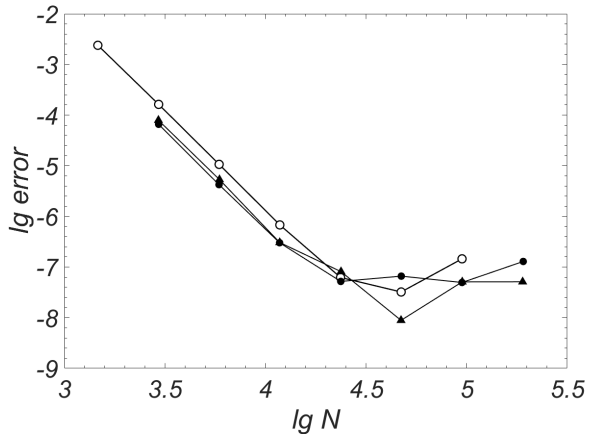

Рис. 2.8. Погрешности при $\lambda = 10^3$ и разных способах вычисления кривизны: ○ – точное выражение (2.35), ▲ – формула (2.18) точности $O(h)$, ● – формула (2.18) точности $O(h^2)$.

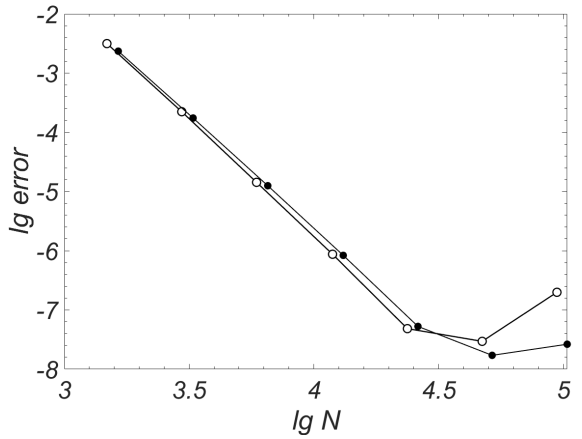

Рис. 2.9. Погрешности при $\lambda = 10^3$; ○ – схема ERK4, ● – BORK4.

ет на точность самого численного решения. Поэтому для вычисления кривизны вполне достаточно формул первого порядка точности, не говоря уже о втором.

2.3.6. Неявные схемы

Одной из наиболее надежных схем четвертого порядка точности является чисто неявная оптимальная схема BORK [60], [61] из класса обратных схем Рунге-Кутты. В этой схеме кривизна рассчитывается по матрице Якоби численного решения, то есть она полностью согласована с численным решением. Напомним, что схема BORK требует итерационного уравнивания на новом слое и несравненно более трудоемка, чем явные схемы.

Сравним расчеты по этой схеме с результатами по явной схеме ERK4, в которой кривизна на первом этапе берется по простейшим формулам первого порядка точности (2.17). На рис. 2.9 приведены графики погрешности второго этапа для этих схем в норме L_2 для достаточно жесткой задачи с $\lambda = 10^3$. Эти кривые почти совпадают. Это показывает, что даже на задачах высокой

жесткости схема ERK4 с выбором шага по кривизне сопоставима по точности и надежности с чисто неявными схемами, хотя ее трудоемкость несравненно меньше.

2.3.7. Известные алгоритмы выбора шага

Наиболее популярными являются следующие пакеты [13]

1. Схема DOPRI5 широко известна. Это явная схема типа Рунге-Кутты. Она имеет порядок точности $p = 5$, и в ней присутствует традиционный алгоритм выбора шага, основанный на вложенных схемах. Хотя эта схема предназначена для нежестких задач, ее нередко применяют для задач умеренной жесткости.
2. Программный пакет Гира (GEAR), предназначенный для жестких задач. Он содержит набор схем с порядками точности с $p = 1$ до $p = 5$, которые имеют устойчивость $L_{1/p}$. Пакет основан на локальном сгущении сеток. Кроме того, он снабжен некоторой автоматикой, выбирающей порядок точности. Он включен в библиотеку MatLab [32] и широко известен.

Эти методы работоспособны на мягких задачах и позволяют получить качественно разумное решение уже при малом числе шагов. При этом вопрос о его количественной точности остается открытым. На более сложных жёстких задачах стандартные программы могут давать сбои вплоть до аварийного останова (т.е. расчет не удается выполнить до конца).

Вычисления производились в среде MATLAB с использованием встроенных реализаций пакетов GEAR и DOPRI5 (функции MATLAB `ode15s` и `ode45` соответственно). Эти реализации широко применяются для практических расчётов, поэтому и используются в данной работе. Для каждого пакета проводилась серия расчётов с уменьшением заданной относительной точности `RelTol`. При заданных значениях `RelTol` меньше $2.22045 \cdot 10^{-14}$ эти функции принудительно увеличивают `RelTol` до $2.22045 \cdot 10^{-14}$.

Срывы автоматики. Традиционные автоматы выбора шага качественно работают следующим образом. При входе в пограничный слой они выбирают достаточно малые шаги $\tau \sim 1/\lambda$. По мере выхода из пограничного слоя они укрупняют шаг и регулярные участки решения считают со сравнительно крупным шагом. Однако именно на регулярных участках наблюдаются срывы шага: последний без видимых причин внезапно уменьшается на 2-4 порядка. После этого автомат снова увеличивает шаг, но срыв шага может повториться, причем неоднократно. Это явление описано [13], однако его причина не была объяснена.

Данное явление было исследовано А.А. Болтневым и О.А. Качер в 2005 году. Они обнаружили, что срывы связаны с процессом медленного увеличения шага на регулярных участках. Если вести расчеты регулярных участков равномерным шагом или увеличивать шаг в геометрической прогрессии со знаменателем, очень близким к 1 ($\sim$1.001), то срывов не происходит. Однако это не позволяет увеличить шаг до нужных пределов за разумное число шагов. Если же увеличивать шаг с разумным значением знаменателя $\sim$1.1, то срывы возникают довольно часто. Поэтому ситуация напоминает известную поговорку “хвост вытащишь – голова увязнет”.

К сожалению, эта работа Болтнева и Качер не была опубликована.

Описанное явление напоминает потерю устойчивости схемами Гершфельдера-Кертисса (на которых построены программы Гира), если вести расчеты по ним не с постоянным шагом, а с увеличением шага в геометрической прогрессии. Потеря устойчивости происходит при тем меньшем знаменателе, чем выше порядок точности схемы. У схемы порядка точности $O(\tau)$ допустимый знаменатель превышает 2, а у схемы $O(\tau^5)$ он составляет $\sim$1.04.

Контроль точности. Известно, что методы вложенных схем и локального сгущения не дают гарантированной оценки точности, то есть фактическая погрешность может отличаться от запрошенной пользователем (tolerance). Проще всего проиллюстрировать это на задаче с известным точным решением. Были проведены расчеты теста (2.31) с $\lambda_0 = 10^5$ по программам GEAR и DOPRI5.

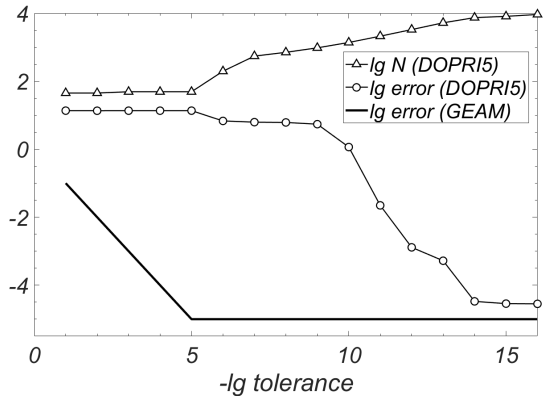

Рис. 2.10. Расчет теста (2.31) для $\lambda_0 = 10^5$. Маркеры – программа DOPRI5: $\circ$ – фактическая погрешность, $\triangle$ – число шагов сетки. Сплошная линия – погрешность ERK4 на геометрически-адаптивных сетках.

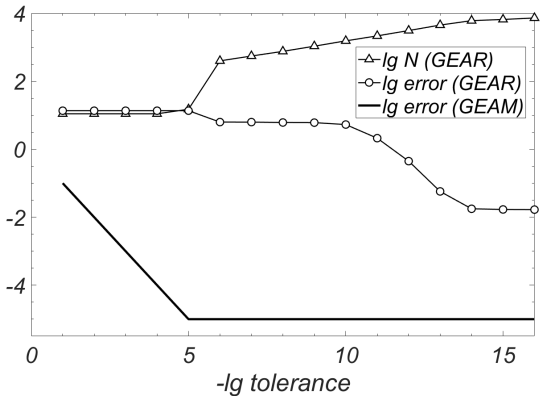

Рис. 2.11. Расчет теста (2.31) для $\lambda_0 = 10^5$. Маркеры – программа Гира. Обозначения маркеров соответствуют рис. 2.10

На рис. 2.10 приведено зависимость логарифма истинной погрешности $\lg \varepsilon$ от логарифма заданной пользователем точности $\lg \mathrm{tolerance}$. Для нашего метода выбора шага и схемы ERK4 $\varepsilon = \mathrm{tolerance}$ вплоть до выхода на ошибки округления (жирная линия). Для программы DOPRI5 величина ε составляет ~ 10 при $\mathrm{tolerance} = 10^{-1} - 10^{-9}$. Заметное уменьшение ε имеет место лишь при значениях $\mathrm{tolerance} < 10^{-10}$, и только при $\mathrm{tolerance} \sim 10^{-14}$ погрешность ε достигает фона ошибок округления $\sim 10^{-5}$. Таким образом, фактическая погрешность ε превышает tolerance на 9-10 порядков в области теоретической сходимости!

На том же графике приведен логарифм числа шагов сетки $\lg N$ в программе DOPRI5 в зависимости от tolerance. Видно, что при tolerance от 10^{-1} до 10^{-5} число шагов составляет ~ 50 и практически не изменяется. При дальнейшем уменьшении tolerance до 10^{-16} число шагов плавно возрастает до 10^4.

Аналогичные результаты имеют место для программы GEAR (см. рис. 2.11). Для нее фон ошибок округления оказывается $\sim 10^{-2}$, то есть выше, чем для

программы DOPRI5. Это означает, что программа Гира в данной задаче в принципе не может дать точность лучше $\sim 10^{-2}$.

Таким образом, чтобы в программах DOPRI5 и GEAR получить разумную точность ε, приходится выбирать неправдоподобно малое tolerance (вплоть до равного ошибке единичного округления). Кроме того, алгоритмы, лежащие в основе этих программ, не очень надежны, и в них сильно сказываются ошибки округления. Величина фона ошибок округления зависит от конкретной задачи и заранее неизвестна. Поэтому возможна такая ситуация, когда требуемая точность оказывается меньше, чем фон ошибок округления, и достигнуть ее в принципе невозможно.

Проведенные расчеты убедительно показывают преимущества предложенного выше способа выбора шага по кривизне и модификации метода Ричардсона для оценки точности.

2.4. О выборе метода при решении жестких задач Коши

Для верификации численного расчета принципиально важно строить апостериорные оценки точности. Однако при любой процедуре сгущения шага апостериорная оценка не может быть меньше ошибок округления. Эти ошибки округления зависят от разрядности чисел компьютера и от характера решаемой задачи. В задачах с начальным пограничным слоем ошибки округления обычно невелики. Однако в задачах с контрастными структурами ошибки округления существенно возрастают. Каждая контрастная структура как бы вычеркивает несколько значащих цифр из численного решения; то есть чем большее число контрастных структур необходимо рассчитать, тем большее число знаков мы потеряем к конечному моменту. Число знаков, которое теряется на каждой контрастной структуре, может существенно зависеть от скорости, с которой решение выходит от пограничного слоя на регулярный участок.

Увеличение разрядности чисел уменьшает влияние ошибок округления. Однако нетрудно построить такие задачи высокой жесткости, на которых никакое разумно приемлемое число разрядов не спасет. Поэтому сеточный расчет становится практически невозможным. Но это как раз тот случай малого параметра $\varepsilon \sim \|\mathbf{f}\|^{-1}$, когда хороший результат могут дать аналитические методы разложения. При этом аналитические методы успешно справляются с любым количеством контрастных структур, что является их дополнительным преимуществом перед сеточными методами.

В то же время на сложных прикладных задачах с большим числом компонент аналитические методы сталкиваются с серьезными трудностями. Примером такой задачи может быть рассмотренный выше расчет горения водорода в кислороде. Эта задача является одной из простейших. Она содержит всего 9 уравнений. Задачи горения топлив могут содержать десятки уравнений, включающих сотни химических реакций. Очевидно, аналитическими методами даже выявить наличие контрастных структур крайне трудно, а построить аналитическое приближение разумной точности практически невозможно. Численные же методы неплохо справляются с такими задачами.

2.4.1. Экспоненциальный тест

Приведем тест, в котором эффект прилипания регулярного решения к стационарному будет определяющим. Возьмем уравнение

$$\frac{du}{dt} = -\lambda(t)u(u^2 - a^2), \quad \lambda(t) = \lambda_0 \cos t. \tag{2.36}$$

У него есть три стационарных решения $u = 0, \pm a$. Контрастные структуры возникают при $\left|u^0\right| < a$. Точное решение задачи имеет вид

$$u(t) = \frac{au^0}{\sqrt{(u^0)^2 + (a^2 - (u^0)^2)\exp\{-2a^2\Lambda(t)\}}}, \quad \Lambda(t) = \int_0^t \lambda(\xi)d\xi. \tag{2.37}$$

При $\lambda(t) > 0$ устойчивыми являются стационары $u = \pm a$, а при $\lambda(t) < 0$ – стационар $u = 0$. Поэтому при “переключении” знака $\lambda(t)$ точное решение по-

переменно притягивается к стационарам $u = a$ (либо $-a$) и $u = 0$. При выборе $\lambda(t) = \lambda_0 \cos t$, $\Lambda(t)$ из (2.32) и $|u^0| < a$ качественный вид решения похож на рис. 2.3. Однако отличие решения от стационара при $|\Lambda(t)| \gg 1$ составляет $\sim \exp\{-2a^2\Lambda(t)\}$. Поэтому скорость налипания экспоненциальная, и сам тест будем называть экспоненциальным. Здесь налипание с точностью до ошибок округления достигается уже при умеренных значениях λ_0. Для расчетов выберем $a = \pi$ и $u^0 = 0.5$.

2.4.2. Численный расчет

Для расчета этой задачи использовалась четырехстадийная оптимальная обратная схема Рунге-Кутты с точностью $O(\Delta t^4)$ и L_4-устойчивостью [59–61]. Эта схема является одной из самых надежных.

В качестве аргумента выбиралась длина дуги. Напомним, что при этом регулярные участки остаются почти горизонтальными, пограничные слои превращаются в почти прямые линии с наклоном ± 1, а переходные зоны по-прежнему имеют большую кривизну.

Расчеты проводились по пакету GEABORK [59] с использованием геометрически-адаптивных сеток. Расчеты выполнялись с 64-разрядными числами. Одновременно с решением вычислялась апостериорная асимптотически точная оценка его погрешности. Полученное решение сравнивалось с точным, а их разница – с апостериорной оценкой погрешности.

Поскольку при расчетах в длине дуги определяются две функции $u(l)$ и $t(l)$, то для сравнения с точным решением мы отсюда вычисляли $u(t)$. Расчеты проводились для трех значений $\lambda_0 = 100,\ 200,\ 400$.

Зависимость погрешности u в норме L_2 от числа узлов расчетной сетки N показана на рис. 2.12. Видно, что при $\lambda_0 = 100$ только начальный участок до $N \approx 250$ вообще не показывает сходимости. Численное решение в какой-то момент прилипает к стационарному решению и далее идет по нему, вообще не описывая последующие контрастные структуры. При дальнейшем увеличении

N виден небольшой участок сходимости, и при $N \approx 1200$ расчет выходит на фон ошибок округления $\sim 10^{-2}$ (то есть теряются 14 знаков из 16!).

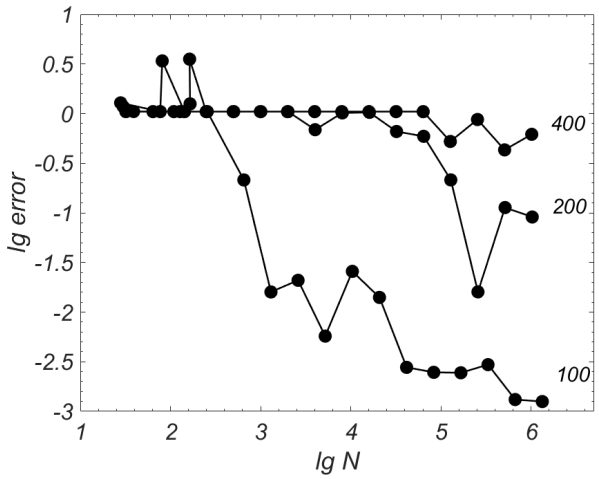

Рис. 2.12. Погрешность в тесте (2.36), цифры около линий – значения λ_0.

При $\lambda_0 = 200$ фон ошибок округления еще выше: $\sim 10^{-1}$. Выход на него происходит на гораздо более подробной сетке с $N \approx 2.5 \cdot 10^5$. При $\lambda_0 = 400$ в пределах наших сеток вообще не видно сходимости: начальный участок непосредственно переходит в фон ошибок округления. Все это показывает, что контрастные структуры с быстрым налипанием на стационарное решение исключительно трудны для сеточных расчетов. Даже переход на расчеты с повышенной разрядностью не позволяет радикально исправить положение.

Заметим, что в этом тесте рассчитать апостериорные оценки погрешности практически невозможно. Нанесенные на рис. 2.12 погрешности определялись как разность численного и точного решений.

2.4.3. Приближенное аналитическое решение

Построение асимптотики. Обычно в теории асимптотических методов рассматриваются контрастные структуры, возникающие в краевых задачах. Для них развита стандартная техника, описанная в [236]. Контрастным структурам в начальных задачах, обусловленным сменой устойчивости, уделяется меньшее внимание. В частности, нам не удалось найти работ, в которых строи-

лись бы асимптотические ряды для решений задач, похожих на тесты (2.31) или (2.36). Поэтому мы построили такие разложения для задач следующего вида:

$$du/dt = -\lambda(t)\varphi(u). \tag{2.38}$$

Очевидно, тест (2.36) являются частными случаями задачи (2.38).

Пусть $\lambda(t)$ достигает экстремумов в точках t_k. Тогда в этих точках производная решения максимальна. Поэтому участки решения в окрестности точек t_k целесообразно считать контрастными структурами. Пусть вырожденное уравнение $\phi(u) = 0$ имеет непересекающиеся однократные корни η_1 и η_2; при $t_{k-1} < t < t_k$ устойчивым является корень η_1, а при $t_k < t < t_{k+1}$ – корень η_2. Тогда в момент t_k возникает контрастная структура, составленная из пограничных слоев слева и справа от t_k.

Эти пограничные слои можно рассматривать как начальные, если взять t_k за начальный момент времени и задать начальные условия

$$\begin{aligned} Q^{\text{лев}} &= u^0 - \eta_1 \quad \text{для левого погранслоя,} \\ Q^{\text{прав}} &= u^0 - \eta_2 \quad \text{для правого погранслоя.} \end{aligned} \tag{2.39}$$

Тогда правый пограничный слой как бы соответствует “движению вперед”: $t \geq 0$, решение исходит из точки (t_k, u^0) и прилипает к η_2. Поэтому его можно рассчитать как обычный начальный пограничный слой стандартными методами [236]. Левый пограничный слой соответствует “движению назад”: $t \leq 0$, решение исходит из точки (t_k, u^0) и прилипает к η_1. Для его расчета нужно в (2.38) сделать замену $t \to -t$ и снова применить стандартный подход.

Заметим, что такой способ приближенного аналитического расчета контрастных структур не является универсальным, так как он применим только к задачам вида (2.38). Требуется также знать моменты t_k. Кроме того, данный способ является эвристическим. Мы не ставим цель его формального обоснования.

Решение экспоненциального теста. Применим описанный выше способ к экспоненциальному тесту (2.36) с $\lambda(t) = \lambda_0 \cos t$. Напомним, что в этом случае $t_k = \pi k$. Для простоты ограничимся первыми членами разложения. Представим

решение в виде $u = \bar{u}_0 + \Pi_0 + Q_0^{\text{лев}} + Q_0^{\text{прав}}$. Здесь $\bar{u}_0$ – регулярное решение, Π_0 – начальный пограничный слой, $Q_0^{\text{лев}}$ и $Q_0^{\text{прав}}$ – левая и правая половины первой контрастной структуры соответственно. Для второй и последующих контрастных структур выражения будут полностью аналогичны.

При $0 < u^0 < a$ устойчивое регулярное решение представляет из себя чередующиеся значения a и 0:

$$\bar{u}_0 = \begin{cases} a, & 0 \le t < \pi,\ 2\pi \le t < 3\pi... \\ 0, & \pi \le t < 2\pi,\ 3\pi \le t < 4\pi... \end{cases} \tag{2.40}$$

Задача для Π_0 строится по стандартной методике и имеет вид

$$\begin{aligned} &\frac{d\Pi_0}{d\tau} = -(a + \Pi_0)[(a + \Pi_0)^2 - a^2], \\ &\Pi_0(0) = u^0 - a. \end{aligned} \tag{2.41}$$

Величина $\tau = t\lambda_0$ называется растянутой переменной. Решение этой задачи имеет структуру, близкую к (2.32):

$$\Pi_0(\tau) = \frac{au^0}{\sqrt{(u^0)^2 + (a^2 - (u^0)^2)\exp\{-2a^2\tau\}}} - a. \tag{2.42}$$

Заметим, что для задачи с $\lambda(t) = \lambda_0 = const$, имеющей только начальный пограничный слой, $\bar{u}_0 + \Pi_0$ является точным решением. Поэтому добавление следующих членов асимптотического разложения заведомо ухудшает точность.

Задача для $Q_0^{\text{лев}}$ $(t < \pi)$ аналогична (2.41), однако теперь растянутая переменная $\tau = -\lambda_0(\pi - t) < 0$. Решение для $Q_0^{\text{лев}}$ имеет вид (2.42), где нужно заменить $\lambda_0 t \to \tau = -\lambda_0(\pi - t)$. Таким образом, в промежутке $0 \le t \le \pi$ решение экспоненциального теста содержит начальный пограничный слой и первую половину первой контрастной структуры и представляется в виде $u = a + \Pi_0 + Q_0^{\text{лев}}$.

Задача для $Q_0^{\text{прав}}$ $(t > \pi)$ также аналогична (2.41). Здесь за растянутую переменную следует взять $\tau = \lambda_0(t - \pi) > 0$, начальное условие $Q_0^{\text{прав}}(0) = u^0$ и подставить их в (2.42). Кроме того, к решению $u = 0 + Q_0^{\text{прав}}$ нужно прибавить первую половину второй контрастной структуры, которая строится аналогично $Q_0^{\text{лев}}$, но теперь относится к моменту $t_2 = 2\pi$.

Точность асимптотики. Сравним полученное аналитическое решение с точным решением (2.37). На рис. 2.13 показан десятичный логарифм погрешности $\lg|\delta u|$ в зависимости от t при намеренно маленьком $\lambda_0 = 10$. На тех же осях отложено точное решение u, десятичный логарифм его производной $\lg|f|$ и кривизны $\lg|\kappa|$. Координату максимума производной целесообразно принять за центр пограничного слоя, а координату максимума кривизны – за центр переходной зоны. Погрешность достигает максимального значения $\sim 10^{-2}$ примерно посередине между этими двумя точками. Внутри пограничного слоя погрешность быстро убывает до нулевого значения в его центре. На регулярном решении погрешность также принимает очень малые значения $\sim 10^{-7} - 10^{-12}$. При увеличении λ_0 структура погрешности остается такой же.

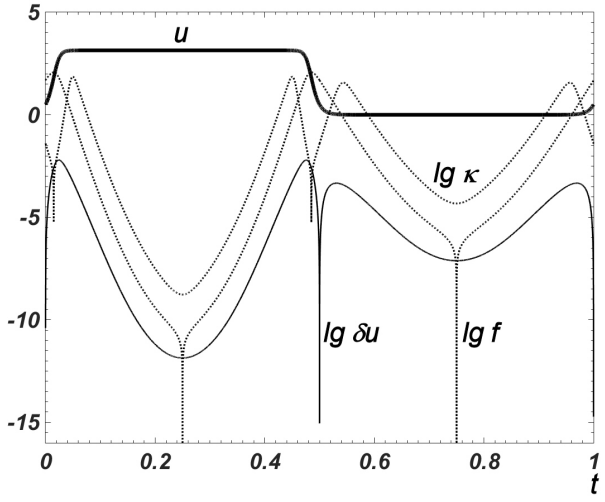

Рис. 2.13. Структура погрешности нулевого приближения. Экспоненциальный тест (2.36), $\lambda_0 = 10$.

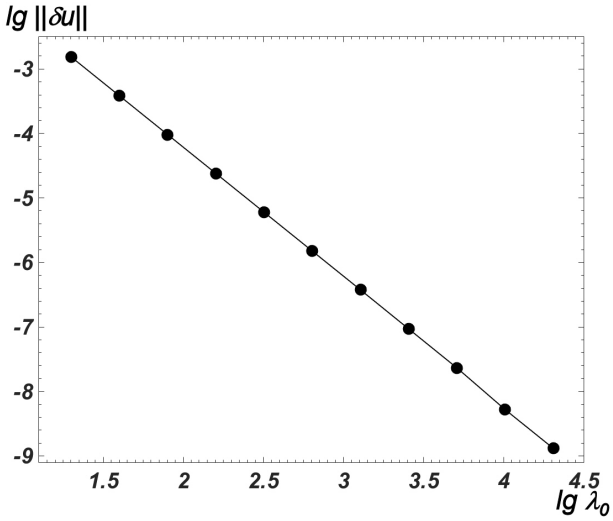

Рис. 2.14. Экспоненциальный тест. Погрешность нулевого приближения.

На рис. 2.14 представлена зависимость погрешности в норме L_2 от величины λ_0 в двойном логарифмическом масштабе. Эта линия является прямой с наклоном 2. Это значит, что остаточный член построенной асимптотики есть величина $O\left(1/\lambda_0^2\right)$. Видно также, что при $\lambda_0 = 100$ (которое бралось в численном расчете) погрешность составляет всего $\sim 5 \cdot 10^{-5}$, что существенно превосходит точность численного расчета. При $\lambda_0 = 400$ погрешность асимптотики равна $\sim 3 \cdot 10^{-6}$, в то время как провести численный расчет и вовсе не удавалось.

Отсюда видны преимущества приближенных аналитических методов перед сеточными на задачах с быстрым прилипанием регулярного решения к стационарному. Приближенные аналитические методы дают очень хорошую погрешность в пограничном слое и переходной зоне и отличную на регулярном решении. Они особенно эффективны при сверхвысокой жесткости и позволяют рассчитать произвольное количество контрастных структур.

2.4.4. Выводы

Сказанное выше показывает, что задачи с контрастными структурами нельзя решить одним универсальным методом. Численные методы достаточно просты, но применимы лишь для не слишком высокой жесткости. Начальный пограничный слой легко рассчитывается ими даже при сверхвысокой жесткости; контрастные же структуры поддаются численному расчету в том случае, когда регулярное решение хорошо отличимо от вырожденого в пределах разрядности используемого компьютера. Однако непонятно, как диагностировать такую ситуацию в ходе численного расчета. Поэтому прежде, чем приступать к решению задачи, следует провести ее качественный анализ. Если жесткость оказывается чрезмерной для данной разрядности компьютера, то следует применять аналитические асимптотические методы.

Наконец, остаются настолько сложные прикладные задачи (например, описанная далее задача химической кинетики), что применение аналитических методов не представляется возможным. Однако обычно эти задачи не являются ультражесткими для разрядности современных компьютеров. Поэтому применение численных методов к ним позволяет (хотя и с трудом) добиться успеха.

Сгущение сеток по времени в случае жестких задач требует неприемлемо большого числа шагов для получения разумной точности. Переход к длине дуги позволяет улучшить расчет пограничных слоев, но этого также недостаточно. Лишь выбор шага с учетом кривизны интегральных кривых позволяет обеспечить высокую точность при умеренном числе шагов и построить хорошие

программы автоматического выбора шага. Тестирование на модельных и реальных задачах показывает, что фактическая точность известных стандартных программ может очень сильно отличаться от требуемой пользователем.

2.5. Основные результаты главы

1. Для определения близости решений жестких задач предложено использовать метрику Хаусдорфа. Построено обобщение этой метрики, аналогичное норме L_2. Показано, что для кривых с контрастными структурами такое определение метрики более адекватно, чем традиционные нормы разности.
2. Для численного решения задачи Коши для ОДУ предложен новый метод автоматического выбора шага по кривизне интегральной кривой. Такие сетки были названы геометрически-адаптивными. Показано, что зависимость шага этих сеток от кривизны является асимптотически оптимальной.
3. Построен способ сгущения геометрически-адаптивных сеток. Он позволяет вычислять апостериорную асимптотически точную оценку погрешности. Известные подходы не дают таких оценок. Поэтому предложенные методы существенно превосходят по надежности известные алгоритмы.
4. Для расчетов по явным схемам построены экономичные методы вычисления кривизны. Для явных схем Рунге-Кутты с числом стадий до 4 приведены таблицы коэффициентов этих формул. Это позволяет успешно применять явные схемы даже к сверхжестким задачам.
5. Проведена апробация предложенных методов на представительных тестовых задачах. Показано, что явные схемы на геометрически-адаптивных сетках не уступают неявным схемам в надежности и точности, но кардинально превосходят их в экономичности. Проведены расчеты тех же тестовых задач по стандартным программам Гира и Дормана-Принса. Эти расчеты показали, что фактическая точность этих программ на много порядков отличается от заданной. Это иллюстрирует преимущества предложенных методов.

6. Проанализированы достоинства и недостатки численных методов и асимптотических методов разложения по малому параметру. Определены области их применимости и даны практические рекомендации. Приведены примеры расчетов, иллюстрирующих эти выводы.

3. Кинетика химических реакций

3.1. Постановка задачи

3.1.1. Система уравнений

Приращение концентрации продукта реакции пропорционально произведению концентраций реагентов. Расход реагента пропорционален произведению его концентрации на концентрацию других реагентов. Коэффициент пропорциональности называется константой скорости элементарной реакции. Он зависит от температуры.

Чтобы найти суммарную скорость изменения j-й концентрации, нужно просуммировать вклады от всех реакций, в которых участвует эта компонента. Если она является продуктом, то соответствующее слагаемое имеет знак «+», если реагентом – то «−». Поэтому система уравнений кинетики реакций имеет следующий вид:

$$\frac{du_j}{dt} = f_j\left(\mathbf{u}, T\right), \;\; f_j\left(\mathbf{u}, T\right) = \sum_{i,l,q=1}^{J} \left(\pm K_{jil}(T) u_i u_l \pm K_{jilq}(T) u_i u_l u_q\right) \qquad (3.1)$$

Здесь u_j – концентрации, J – число компонент.

Газофазные реакции могут протекать как в прямом, так и в обратном направлении. Под прямой реакцией будем понимать экзотермическую (то есть протекающую в направлении выделения тепла), а под обратной – эндотермическую (в которой тепло поглощается). Выделяемая или поглощаемая энергия E называется энергией реакции. При составлении системы (3.1) каждое направление следует считать за отдельную реакцию.

3.1.2. Особенности задачи

Задачи кинетики реакций имеют ряд специфических особенностей. Они важны для построения специализированных схем.

1° Концентрации u_j являются неотрицательными.

2° Правые части каждого уравнения разбиваются на положительные и отрицательные слагаемые. Положительные соответствуют тем реакциям, в которых j-ое вещество образуется, а отрицательные – реакциям, в которых это вещество расходуется. Скорости реакций с расходом пропорциональны u_j или u_j^2, так что в них всегда можно явно выделить множитель u_j. Таким образом, задачу (3.1) можно записать в виде

$$\frac{du_j}{dt} = f_j\left(\mathbf{u}\right), \quad f_j(\mathbf{u}) = -u_j\varphi_j(\mathbf{u}) + \psi_j(\mathbf{u}), \tag{3.2}$$

причем $u_j \geqslant 0$, $\varphi_j(\mathbf{u}) \geqslant 0$, $\psi_j(\mathbf{u}) \geqslant 0$.

3° При химических реакциях меняется число молекул, но сохраняются числа атомов каждого элемента. Математически это означает наличие линейных первых интегралов системы дифференциальных уравнений. Пусть имеется такой набор констант α_j, что $\sum \alpha_j f_j = 0$. Тогда у точного решения существует первый интеграл $\sum \alpha_j u_j(t) = \text{const}$. Это сохранение одного химического элемента. Каждому химическому элементу соответствует свой набор констант α_j и свой первый интеграл. Пусть j-я молекула содержит γ_{ji} атомов сорта i. Тогда должно сохраняться полное число атомов сорта i, которое равно

$$\sum_{j=1}^{J} \gamma_{ji} u_j, \quad 1 \leqslant i \leqslant I. \tag{3.3}$$

Здесь I – число химических элементов.

В [59, 237] доказано, что линейные первые интегралы точно передаются в схемах Рунге-Кутты и Розенброка. Нарушение этих балансов может возникать только за счет ошибок машинного округления.

Построению численных схем, сохраняющих различные интегралы движения, посвящена обширная литература (см., например, [165, 166, 168–174, 238] и библиографию там). Подробное исследование интегралов задачи (3.1) выходит за рамки данной работы.

Заметим, что после перехода к длине дуги указанные выше особенности задачи кинетики остаются в силе. В частности, правые части новой системы

также априори разделяются на положительные и отрицательные слагаемые. Они имеют вид $F_j = -U_j\Phi_j + \Psi_j$, где

$$\Phi_0 = 0, \quad \Psi_0 = \frac{1}{S}, \quad \Phi_j = \frac{\nu_0\varphi_j}{S}, \quad \Psi_j = \frac{\nu_0\psi_j}{S}, \quad 1 \leqslant j \leqslant J. \tag{3.4}$$

3.1.3. Трудности

Сложность задачи обусловлена следующими факторами. ***1°*** Число компонент обычно велико. Например, горение водорода в кислороде – это простейшая система реакций, учитывающая около десятка веществ. При горении водорода в воздухе нужно учитывать ~ 30 соединений с числом атомов до пяти. Горение метана в воздухе – это еще более сложный процесс, в котором число компонент может достигать нескольких десятков, а число реакций – нескольких сотен.

2° Скорости различных реакций (с учетом прямых и обратных) могут отличаться на 10 и более порядков. Поэтому задача кинетики реакций является жесткой. Это справедливо для реакций различной природы, как химических, так и ядерных.

3° Есть еще одна принципиальная трудность, присущая химическим и плазмохимическим реакциям. К реагирующим веществам могут быть добавлены небольшие присадки – ингибиторы или катализаторы. Пусть в системе учтены ингибиторы и все реакции с их участием. Если начальные концентрации присадок равны нулю, то реакции протекают по обычному механизму. Если начальные концентрации ингибиторов малы, но отличны от нуля, то горение существенно замедляется или вовсе затухает. Напротив, небольшая добавка катализатора ускоряет реакцию.

Задачи, в которых решение сильно зависит от входных данных, называются плохо обусловленными. Задача кинетики реакций очень плохо обусловлена. Это является серьезной трудностью для всех численных методов. В плохо обусловленных задачах ошибки округления оказываются очень велики.

3.2. Химические схемы

3.2.1. Специальные схемы

Специфические особенности задачи позволяют построить явную схему, обладающую малой трудоемкостью.

Введем обозначения: t – исходный момент времени, $\hat{t} = t + \tau$ – новый момент времени, u_j и $\hat{u}_j$ – решения в эти моменты. Калиткин и Гольдин предложили [239] специальную явную схему точности $O(\tau)$. Решение в новый момент времени по этой схеме обозначим через $\tilde{u}_j$; оно равно

$$\tilde{u}_j = \frac{u_j + \tau\psi_j(\mathbf{u})}{1 + \tau\varphi_j(\mathbf{u})}. \tag{3.5}$$

Эта схема разумна: увеличение $\psi_j(\mathbf{u})$ как в точном, так и в численном решении приводит к увеличению образования вещества; увеличение $\varphi_j(\mathbf{u})$ действует противоположно. При этом u_j остается неотрицательным.

Недостатком схемы (3.5) является невысокая точность. В [239] предлагались методы повышения порядка точности до второго. Однако они оказались неудачными: в численном решении возникала сильная немонотонность, из-за чего порядок точности фактически не повышался. Кроме того, схема теряла надежность; например, начальные концентрации нельзя было задавать нулями, нужно было вводить малые числа.

В данной работе построена явная схема второго порядка точности, одновременно имеющая более высокую надежность. Приведем ее. Напишем следующую неявную схему:

$$\hat{u}_j = \frac{u_j + \tau\psi_j(\overline{\mathbf{u}})(1 + \tau\varphi_j(\overline{\mathbf{u}})/2)}{1 + \tau\varphi_j(\overline{\mathbf{u}}) + (\tau\varphi_j(\overline{\mathbf{u}}))^2/2}, \quad \overline{\mathbf{u}} = (\mathbf{u} + \hat{\mathbf{u}})/2. \tag{3.6}$$

Будем находить решение алгебраической системы простыми итерациями:

$$\hat{u}_j^{s+1} = \frac{u_j + \tau\psi_j(\overline{\mathbf{u}}^s)(1 + \tau\varphi_j(\overline{\mathbf{u}}^s)/2)}{1 + \tau\varphi_j(\overline{\mathbf{u}}^s) + (\tau\varphi_j(\overline{\mathbf{u}}^s))^2/2}, \quad \overline{\mathbf{u}}^s = (\mathbf{u} + \hat{\mathbf{u}}^s)/2, \quad \hat{\mathbf{u}}^0 = \mathbf{u}. \tag{3.7}$$

При этом выполним только две итерации, то есть по существу получим явную схему. Здесь также увеличение ψ_j приводит к увеличению $\hat{u}_j$, а увеличение φ_j –

к уменьшению $\hat{u}_j$. Схемы (3.5) и (3.7) назовем ***химическими схемами*** (одно- и двухстадийной соответственно).

3.2.2. Свойства специальных схем

Аппроксимация и устойчивость. Разложением в ряды можно доказать, что первая итерация схемы (3.7) имеет аппроксимацию $O(\tau)$, а вторая итерация – $O(h^2)$. Третья и последующие итерации не улучшают порядок точности, но повышают трудоемкость схемы. Поэтому выполнять их не следует.

Нетрудно убедиться, что на линейном тесте Далквиста

$$du/dt = -\lambda u \tag{3.8}$$

схема (3.7) является L_2-устойчивой.

Трудоемкость. Схема (3.7) явная, поэтому расчеты по ней имеют малую трудоемкость. В самом деле, расчеты на каждой итерации требуют однократного вычисления J правых частей. Таким образом, трудоемкость двухстадийной химической схемы такова же, как у двухстадийной явной схемы Рунге-Кутты, что гораздо меньше трудоемкости неявных схем. Последние требуют нахождения матрицы Якоби, что соответствует вычислению J^2 правых частей.

Таким образом, схема (3.7) в $\sim J$ раз менее трудоемка, чем явно-неявные схемы и схемы Розенброка и Розенброка-Ваннера. Выигрыш по сравнению с чисто неявными итерационными схемами еще больше. Это преимущество особенно существенно для систем большого порядка, когда в реакциях участвует большое число компонент.

Неконсервативность. Как отмечалось в п. 3.1, в точном решении суммарное число атомов каждого элемента всегда остается постоянным. Такой баланс есть линейный первый интеграл системы (3.2). В методе (3.5) и (3.7) для численного решения эти интегралы передаются не точно, а приближенно с точностью $O(\tau)$ для схемы (3.5) и $O(\tau^2)$ для схемы (3.7).

Это не представляет трудностей при расчете. В задаче (3.1) решение является классическим, обобщенных решений нет. Поэтому достаточно провести расчеты со сгущением сеток. Численные решения, полученные на разных сетках, будут стремиться к предельной функции при $\tau \to 0$. Поскольку решение классическое, явление ложной сходимости отсутствует, и предельная функция будет искомым точным решением. Поэтому нарушение баланса будет стремиться к нулю при $\tau \to 0$. При этом оно имеет приблизительно тот же порядок величины, что и погрешность решения. Поэтому дисбаланс может служить дополнительным средством контроля точности.

Знакопостоянность. Поскольку $u_j \geqslant 0$, $\varphi_j(\mathbf{u}) \geqslant 0$, $\psi_j(\mathbf{u}) \geqslant 0$, то, очевидно, $\hat{u}_j \geqslant 0$ и $\tilde{u}_j \geqslant 0$. Поэтому численное решение по схемам (3.5) и (3.7) является знакопостоянным. Это соответствует физическому смыслу задачи (3.1) и является достоинством этих схем. Далее будет показано, что неявные схемы знакопостоянности не гарантируют.

Немонотонность. Схемы (3.5) и (3.7) являются немонотонными, то есть при монотонном точном решении численное решение может иметь участки немонотонности или осциллировать. Однако на нелинейных жестких задачах неявные схемы также иногда оказываются немонотонными. Для исследования этого свойства мы проводили расчеты специальной тестовой задачи, представленной ниже.

3.2.3. Апробация

Постановка. Рассмотрим задачу для одного уравнения с кубической правой частью, имитирующую химические реакции со столкновением 3 одинаковых частиц:

$$\frac{du}{dt} = -\lambda u \left(u^2 - a^2\right); \quad a > 0, \quad \lambda \gg 1; \quad u(0) = u^0. \tag{3.9}$$

Эта задача удобна тем, что для нее легко построить точное решение

$$u(t) = \frac{au^0}{\sqrt{(u^0)^2 + \left[a^2 - (u^0)^2\right] \exp\{-2\lambda a^2 t\}}}. \tag{3.10}$$

Поле интегральных кривых задачи (3.9) приведено на рис. 3.1. Точное решение имеет три стационара $u(t) = a, 0, -a$; из них первый и третий устойчивые, а второй неустойчивый. Точное решение имеет пограничный слой шириной $t \sim 1/\lambda a^2$ и быстро выходит на 1-й стационар при $u^0 > 0$ и на 3-й при $u^0 < 0$.

Если u имеет смысл концентрации, то осмысленным является только положительное решение $u(t) > 0$, а отрицательные решения физического смысла не имеют.

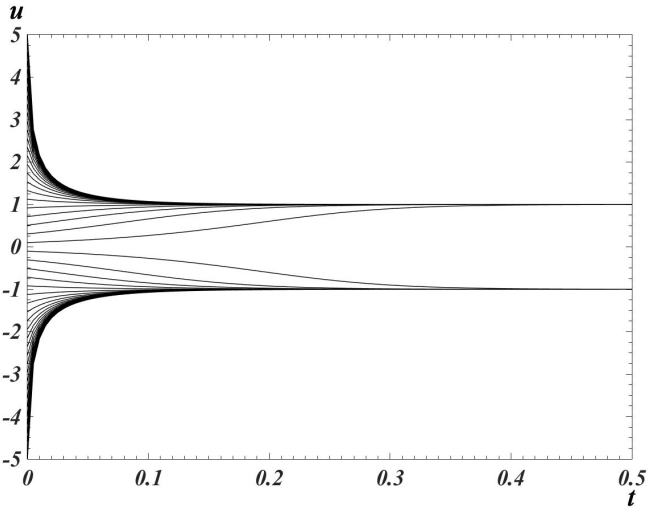

Рис. 3.1. Поле интегральных кривых для теста (3.9) при $a = 1$.

Мы провели расчеты этой задачи по следующим схемам: одно- и двухстадийная химические схемы, чисто неявная схема Розенброка (1.3) с коэффициентом 1, схема Розенброка (1.3) с комплексным коэффициентом $(1 + i)/2$, неявная схема Эйлера (1.4).

Отметим одну деталь. Одно- и двухстадийная химические схемы есть соответственно первая и вторая простые итерации для неявной схемы (3.6). Для этих схем порядок точности равен числу итераций. Чисто неявная схема Розенброка (1.3) есть первая ньютоновская итерация для неявной схемы Эйлера (1.4); она имеет первый порядок точности. Возникает естественное желание

выполнить вторую ньютоновскую итерацию. При этом получается некоторая явно-неявная схема. Однако разложение в ряды по степеням τ показывает, что это не повышает порядка точности. Поэтому такую модификацию схемы мы рассматривать не будем.

Расчет по химическим схемам. Задаче (3.9) соответствуют $\varphi = \lambda u^2$, $\psi = \lambda u a^2$. Проанализируем поведение решения (3.5) вблизи стационара. Эту схему можно преобразовать к виду

$$\hat{u} - a = (u - a)\frac{1 - \tau\lambda a u}{1 + \tau\lambda u^2} \tag{3.11}$$

Если $\tau\lambda a u > 1$ (то есть сетки достаточно грубые), то дробь отрицательна. Тогда $\hat{u} - a$ и $u - a$ имеют разные знаки, и выход на стационар оказывается немонотонным. Решение начинает осциллировать вокруг стационарного значения, причем амплитуда осцилляций убывает со временем. Это не препятствует сходимости (так как амплитуда осцилляций уменьшается как $O(\tau)$ при $\tau \to 0$), но делает неправильным качественное поведение решения (см. рис. 3.2).

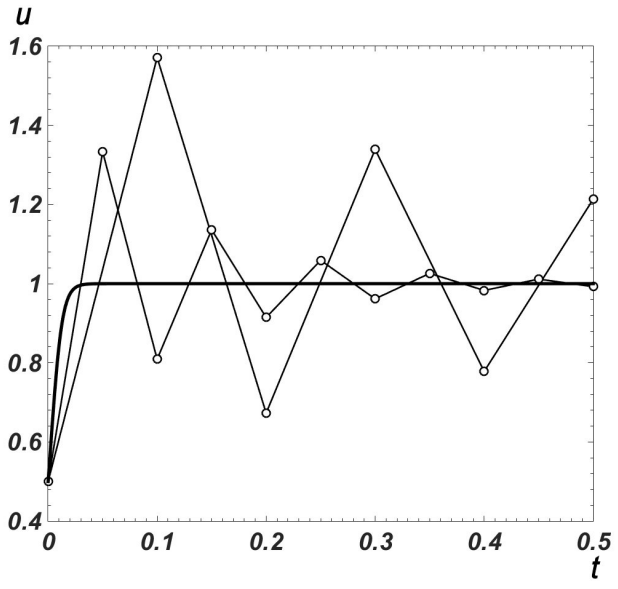

Рис. 3.2. Решение теста (3.9) по одностадийной химической схеме (3.5); ∘ – расчетные точки, жирная линия – точное решение.

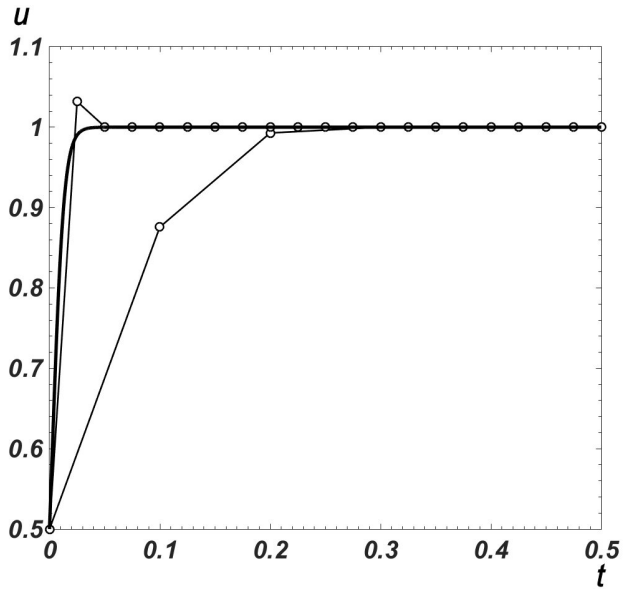

Рис. 3.3. Решение теста (3.9) по двухстадийной химической схеме (3.7); обозначения соответствуют рис. 3.2.

Аналогично проанализируем первую стадию (3.7); для нее $\overline{u} = u$, и

$$u^{(1)} - a = (u - a)\frac{1 - \tau\lambda a u - \tau^2\lambda^2 a u^3/2}{1 + \tau\lambda u^2 + \tau^2\lambda^2 u^4/2}. \tag{3.12}$$

Знаменатель в правой части всегда положителен, а числитель при достаточно большом τ становится отрицательным. Это значит, что $u^{(1)} - a$ и $u - a$ имеют разные знаки, и на первой стадии решение "перепрыгивает" через стационар.

Поэтому если вести расчет по схеме (3.7) только с одной стадией, то численное решение будет иметь пилообразный вид. Результаты расчета с двумя стадиями представлен на рис. 3.3.

Решение на самой грубой сетке с $N = 10$ шагами имеет качественно правильный вид. В этом случае шаг заметно больше ширины пограничного слоя, то есть все узлы сеток лежат в области регулярного решения. На следующей сетке с $N = 40$ первый шаг "перепрыгивает" на другую сторону стационара $u = 1$, и далее решение стремится к этому стационару с неправильной стороны. Здесь шаг близок к ширине пограничного слоя; эти условия наиболее трудны для схемы. Решение на более подробных сетках имеют правильное качественное поведение, поскольку они уже разумно разрешают пограничный слой.

Отметим, что ни на одном из решений мы не видели осцилляций в отличие от одностадийной схемы (3.5). Это означает, что двухстадийная химическая схема (3.7) не только обладает лучшей точностью, но и одновременно является более надежной, чем известная ранее схема (3.5).

Сравнение с неявными схемами. Чисто неявная и комплексная схемы Розенброка, а также неявная схема Эйлера известны как монотонные, так как они дают монотонное решение на линейном тесте Далквиста (3.8). Однако на нелинейном тесте (3.9) картина оказывается иной.

Расчеты теста (3.9) по этим схемам показали, что все эти схемы могут становиться немонотонными (см. рис. 3.4). На грубых сетках решение может и вовсе притягиваться к отрицательному или нулевому стационарам, что противоречит физическому смыслу задачи.

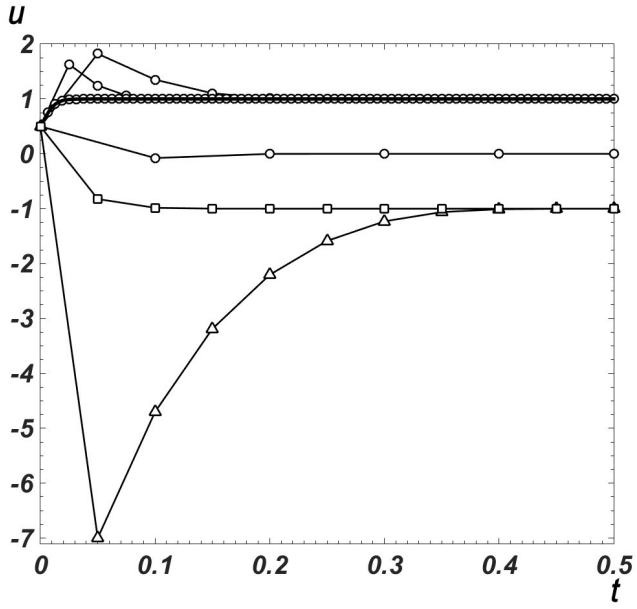

Рис. 3.4. Решение теста (3.9) по неявным схемам: △ – чисто неявная схема Розенброка, ∘ – CROS, □ – неявная схема Эйлера; жирная линия – точное решение.

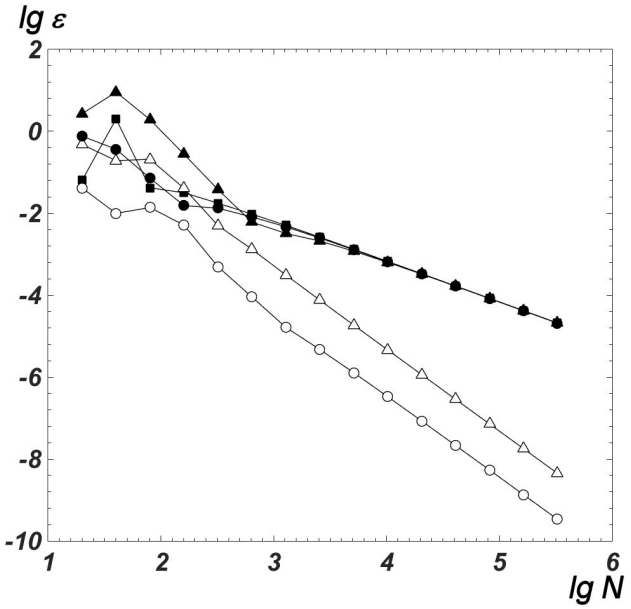

Рис. 3.5. Оценки погрешности по методу Ричардсона в тесте (3.9); ▲ – чисто неявная схема Розенброка, △ – комплексная схема Розенброка, ■ –неявная схема Эйлера, ● – одностадийная химическая схема, ∘ – двухстадийная химическая схема.

Причина этого заключается в следующем. Несмотря на то, что решение исходной дифференциальной задачи (3.9) единственно, нелинейное алгебраическое уравнение относительно $\hat{u}$ может несколько корней. Если задача жесткая, а шаг грубый, то нулевое приближение, выбранное с предыдущего шага, может оказаться неудачным. В результате итерационный процесс может сойтись к неправильному корню.

Таким образом, на задаче кинетики (3.1) неявные схемы не имеют преимуществ перед химическими схемами и при этом более трудоемки.

Расчеты в длине дуги. Метод Ричардсона можно применять как в аргументе время, так и в аргументе длина дуги [237, 240]. В этом случае оценка погрешности по методу Ричардсона дает погрешность $\Delta_j(l)$ (включая $j = 0$, соответствующее $n_0 \equiv t$) как функцию длины дуги, что не всегда удобно на

практике. При химических и термоядерных экспериментах выход продукта регистрируется в определенные моменты времени. К этим моментам и должны относиться оценки погрешности.

В данной работе для расчетов в аргументе длина дуги построена ричардсоновская оценка точности как функции расчетного времени $\delta_j(t)$. Эта оценка имеет вид

$$\delta_j(t) = \Delta_j(l) - f_j(\mathbf{u})\Delta_0(l), \quad 1 \leqslant j \leqslant J. \tag{3.13}$$

Оценку (3.13) назовем ***приведенной***. Из свойств оценок по методу Ричардсона следует, что она асимптотически точна. Эта оценка используется во всех расчетах, приведенных в данной главе.

Оценки погрешности. Проведем расчет на наборе сеток по алгоритму, описанному в п. 1.4. График погрешности в задаче (3.9) представлен на рис. 3.5. На первых нескольких сетках погрешность ведет себя нерегулярно, что говорит о неправильном качественном поведении решений. Однако начиная с достаточно подробных сеток, линии выходят на прямые с наклоном, равным теоретическому порядку точности соответствующей схемы. Видно также, что двухстадийная химическая схема позволяет получать гораздо более высокие точности, чем все схемы первого порядка и несколько более высокие, чем гораздо более трудоемкая схема CROS.

Выводы. Проведенные расчеты показывают, что двухитерационная химическая схема она очень перспективна для задач, сводящихся к (3.2) и, в частности, задач кинетики реакций. Эта схема является более надежной, чем явные схемы Рунге-Кутты. В этом и заключается ее преимущество. Она может быть эффективна в сложных комплексных задачах, когда подгонять шаг по времени только под реакции практически невозможно. Примером является задача течения реагирующих смесей, в которой учитывается не только кинетика реакций, но и газодинамические процессы. В таких задачах следует применять высоконадежные схемы (даже если при этом приходится жертвовать высоким порядком

точности). Химическая схема уже была успешно применена газодинамиками в расчетах течения реагирующих газовых смесей [241].

Задача (3.1) является жесткой, а двухитерационная химическая схема относится к явным. Может показаться, что это противоречит теореме Далквиста. Однако здесь нет никакого противоречия. В определении устойчивости по Далквисту ошибка не должна нарастать при любом шаге. В химических схемах такая устойчивость имеет место в линейном тесте (3.8), но в нелинейных задачах ошибка не нарастает лишь при достаточно малом шаге. Именно это условие обеспечивает нужное поведение вблизи пограничного слоя (детальное исследование этого вопроса выходит за рамки данной работы). Подчеркнем, что на сверхжестких задачах результаты расчетов как по схеме (3.7), так и по другим схемам следует контролировать по визуальной сходимости профилей решения на сгущающихся сетках и по графикам погрешности и дисбаланса.

3.2.4. Пакет программ

Предлагаемый алгоритм реализован в виде прикладного пакета программ `GACK_GEAD` в среде Matlab. Это не самый эффективный язык с точки зрения быстродействия, и для решения прикладных задач данные пакеты невыгодны. Однако в них решены и отлажены все принципиальные алгоритмические вопросы. Кроме того, коды на Matlab легко переносятся на более эффективные языки программирования (в первую очередь Fortran и несколько сложнее C/C++). Поэтому предлагаемые пакеты могут служить хорошими прототипами для высокопроизводительных производственных программ (создание последних выходит за рамки данной работы).

Пакет предназначен для расчетов кинетики реакций на геометрически-адаптивных сетках. Он содержит реализацию двухстадийной химической схемы и алгоритм сгущения сеток и получения решения вместе с апостериорной оценкой погрешности. Отметим, что при замене решателя на общую схему (например, Рунге-Кутты) данный пакет можно применять и к другим задачам.

Пакет `GACK_GEAD` распространяется по свободной лицензии BSD-3-clause и доступен по ссылке `https://github.com/ABelov91/GACK_GEAD`.

3.3. Горение водорода в кислороде

3.3.1. Система реакций

Рассмотрим горение водорода в кислороде. При этом будем учитывать следующие компоненты:

$$O_2,\ H_2,\ O,\ H,\ OH,\ HO_2,\ H_2\ O_2,\ H_2O,\ O_3. \tag{3.14}$$

Между этими компонентами будем учитывать по 25 прямых и столько же обратных реакции, они перечислены в табл. 3.1. Это наиболее полная система реакций, собранная нами по ряду тематических публикаций и авторитетных баз данных [242–245]. Исходными компонентами являются O_2 и H_2. Их концентрации равны $1.5 \cdot 10^{-5}$моль/см^3 и $3 \cdot 10^{-5}$моль/см^3 соответственно. Такие концентрации соответствуют плотности воздуха при комнатной температуре и атмосферном давлении.

Поскольку полное число реакций равно 50, а в каждое уравнение может входить несколько реакций, то эта система может оказаться очень громоздкой, и мы ее не приводим. В качестве примера напишем уравнение для концентрации H. Оно представимо в виде

$$du_4/dt = \varphi_4 - \psi_4, \tag{3.15}$$

где φ_4 соответствует наработке этой компоненты

$$\begin{aligned}\varphi_4 = {} & K_1^r u_2 u_3 + K_2^f u_3 u_5 + K_5^r u_1 u_2 + K_6^r u_5^2 + K_7^r u_3 u_8 + K_8^f u_2 u_5 + K_9^r u_2 u_6 + \\ & + K_{10}^r u_5 u_8 + 2K_{11}^r u_2 \sum_{j=1}^{J} u_j + K_{16}^r u_6 \sum_{j=1}^{J} u_j + K_{17}^r u_8 \sum_{j=1}^{J} u_j + \\ & + K_{19}^r u_5 \sum_{j=1}^{J} u_j + K_{23}^r u_1 u_5,\end{aligned} \tag{3.16}$$

а ψ_4 – расходу

$$\begin{aligned}\psi_4 = {} & K_1^f u_4 u_5 + K_2^r u_1 u_4 + K_5^f u_4 u_6 + K_6^f u_4 u_6 + K_7^f u_4 u_6 + K_8^r u_4 u_8 + \\ & + K_9^f u_4 u_7 + K_{10}^f u_4 u_7 + 2K_{11}^f u_4^2 \sum_{j=1}^{J} u_j + K_{16}^f u_1 u_4 \sum_{j=1}^{J} u_j + \\ & + K_{17}^f u_4 u_8 \sum_{j=1}^{J} u_j + K_{19}^f u_3 u_4 \sum_{j=1}^{J} u_j + K_{23}^f u_4 u_9.\end{aligned} \tag{3.17}$$

Здесь компоненты занумерованы в соответствии с (3.14); K – константы реакций, верхние индексы f и r обозначают прямую и обратную реакцию соответственно, нижний индекс у K соответствует номеру реакции в табл. 3.1 [246].

Таблица 3.1. Реакции горения водорода в кислороде.

Реакция	E, эВ	$\lg C$	Реакция	E, эВ	$\lg C$
$OH+H \rightleftarrows H_2+O$	0.087	12.61	$OH+HO_2 \rightleftarrows H_2O+O_2$	3.074	13.07
$O+OH \rightleftarrows O_2+H$	0.725	13.86	$OH+H_2O_2 \rightleftarrows H_2O+HO_2$	1.272	12.40
$O+HO_2 \rightleftarrows OH+O_2$	2.401	13.29	$2HO_2 \rightleftarrows H_2O_2+O_2$	1.802	11.83
$O+H_2O_2 \rightleftarrows OH+HO_2$	0.599	12.51	$H+O_2+M \rightleftarrows HO_2+M$	1.802	15.37
$H+HO_2 \rightleftarrows H_2+O_2$	2.488	13.45	$H+OH+M \rightleftarrows H_2O+M$	5.067	15.79
$H+HO_2 \rightleftarrows 2OH$	1.676	13.59	$2O+M \rightleftarrows O_2+M$	5.119	14.46
$H+HO_2 \rightleftarrows H_2O+O$	2.349	12.73	$O+H+M \rightleftarrows OH+M$	4.394	15.31
$OH+H_2 \rightleftarrows H+H_2O$	0.586	12.97	$O_2+O+M \rightleftarrows O_3+M$	1.055	14.87
$H+H_2O_2 \rightleftarrows H_2+HO_2$	0.686	12.31	$OH+O_3 \rightleftarrows O_2+HO_2$	1.663	11.32
$H+H_2O_2 \rightleftarrows H_2O+OH$	2.948	11.76	$O_3+O \rightleftarrows 2O_2$	4.064	12.39
$2H+M \rightleftarrows H_2+M$	4.481	14.82	$O_3+H \rightleftarrows OH+O_2$	3.339	12.56
$2OH \rightleftarrows H_2O+O$	0.673	12.77	$O_3+HO_2 \rightleftarrows 2O_2+OH$	1.346	9.69
			$O+H_2O \rightleftarrows H_2+O_2$	0.139	-0.30

Эта химическая задача является одной из простейших. Она содержит всего 9 уравнений. Задачи горения топлив могут содержать десятки уравнений, включающих сотни химических реакций. Очевидно, аналитическими методами даже

выявить наличие контрастных структур крайне трудно, а построить аналитическое приближение разумной точности практически невозможно. Численные же методы неплохо справляются с такими задачами.

Выражения для скоростей реакций возьмем согласно [246]: для экзотермической реакции скорость реакции равна

$$K(T) = C\sqrt{(\pi E/4) + T}. \tag{3.18}$$

Для эндотермической реакции имеем

$$K(T) = C\sqrt{(\pi E/4) + T}\exp\left(-E/T\right). \tag{3.19}$$

Коэффициенты C и E приведены в табл. 3.1

Температура в ходе расчета не изменяется. Такая постановка описывает горение в проточном реакторе, в котором с малой скоростью течет инертный газ (например, аргон), содержащий небольшую примесь водорода и кислорода. В этом случае прогревом смеси за счет энерговыхода реакций можно пренебречь, а все концентрации считать однородными по сечению. Данная постановка широко используется при экспериментальном измерении скоростей химических реакций. Расчеты проводились при двух температурах: умеренной 2000 K и высокой 6000 K.

3.3.2. Профили концентраций

Рассмотрим результаты расчетов. В качестве аргумента выбиралась длина дуги интегральной кривой, сетки были геометрически-адаптивными (если не оговорено обратное). Однако длина дуги не имеет прямого физико-химического смысла: она составляется из таких компонент, как время и концентрации, которые имеют разную размерность. Поэтому результаты представлены как зависимости концентраций и других величин от времени. Это нетрудно сделать, так как каждый узел сетки по l содержит в себе и момент времени, и все концентрации в этот момент.

Температура 2000 K. В первом расчете температура равнялась $T = 2000$ K. Расчеты проводились по трем схемам: ERK2, ERK4 и специальной схеме (3.7). Все схемы давали сходимость к одним и тем же предельным кривым. На рисунке 3.6 приведены предельные кривые концентраций различных видов частиц в зависимости от времени. Видно, что наиболее значимыми компонентами являются O_2, H_2, O, H, H_2O и OH. Концентрации остальных компонент не превышают 3% от полной концентрации смеси.

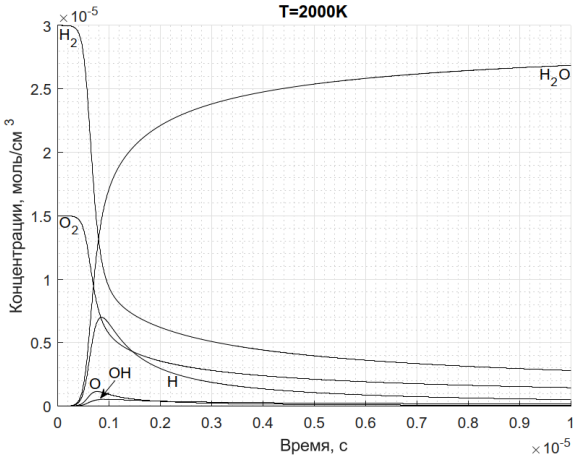

Рис. 3.6. Решение при температуре $T = 2000$ K.

При данной температуре основным продуктом горения является вода. Ее концентрация монотонно нарастает, а концентрации молекулярных кислорода и водорода монотонно убывают.

Заметим, что все концентрации почти постоянны до момента $t \sim 0.5$ мкс, а затем быстро изменяются за характерное время ~ 0.25 мкс. Далее они плавно выходят на стационарные значения к моменту $t \sim 10$ мкс. Это резкое изменение указывает на существование пограничного слоя, причем этот слой является внутренним (то есть имеет место не в начальный момент времени, а внутри промежутка интегрирования). Напомним, что внутренний пограничный слой называют также контрастной структурой.

Поясним причину появления контрастной структуры. В начальный момент присутствуют только чистые H_2 и O_2. Реакции распада этих молекул на атомы имеют высокий энергетический порог, реакции между этими молекулами также

пороговые (то есть все эти реакции обратные). Поэтому сначала реакции протекают медленно. Затем, по мере накопления других компонент оказываются возможными беспороговые реакции с участием H_2 и O_2. Тогда процесс горения существенно ускоряется, и начинается интенсивное горение. Оно соответствует резкому изменению концентраций основных компонент. Далее смесь смесь выгорает, и реакции существенно замедляются.

Влияние компонент на протекание реакций. Из рис. 3.6 видно, что при $T = 2000$ K основными компонентами являются O_2, H_2, O, H, H_2O. Концентрации остальных компонент не превышают 2% от полной концентрации смеси. Однако это не значит, что их можно не учитывать.

Проиллюстрируем это следующим вычислением. Вычеркнем все реакции, содержащие одну из этих компонент, и выполним расчет при тех же условиях ($T = 2000$ K, $t \leqslant 10$ мкс, начальные условия соответствуют гремучей смеси при нормальном давлении).

Компонента OH входит в 15 реакций из 24, а ее максимальная концентрация равна 1.4% от полной концентрации смеси. Однако при отбрасывании этой компоненты горение вообще не идет: концентрации практически не меняются с течением времени. Это значит, что в исходной системе реакций компонента OH активно нарабатывается и расходуется примерно с той же скоростью, и отбрасывать ее недопустимо.

Компонента HO_2 также входит в большое количество реакций (12 из 24), а ее максимальное концентрация составляет 0.4% от полной концентрации смеси. Однако при отбрасывании этой компоненты горение идет и качественное поведение остальных компонент близко к изображенному на рис. 3.6. При отбрасывании компонент O_3 и H_2O_2 картина аналогичная. Это объясняется тем, что при данной температуре эти 3 компоненты не нарабатываются в достаточном количестве и практически не влияют на течение реакций.

Подчеркнем, что при других температурах результаты могут быть другими, и без расчета неочевидно, какие компоненты важны, а какие – нет. Выбор

существенных компонент и отбрасывание лишних представляет собой сложную химическую задачу. Проведенные расчеты показывают, что если система реакций недостаточно полна, то результаты могут оказываться качественно неправильными. Поэтому в практических расчетах рекомендуется «не экономить» на количестве реакций и компонент; кроме того, предложенные методы позволяют легко обрабатывать большие системы реакций.

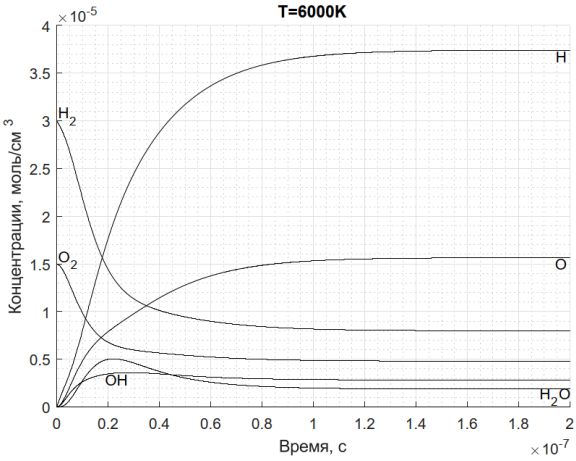

Рис. 3.7. Решение при температуре $T = 6000$ K.

Температура 6000 K. На рисунке 3.7 приведены аналогичные результаты расчетов при температуре $T = 6000$ K. Видно, что при такой температуре концентрации выходят на стационарные гораздо раньше, то есть процесс протекает быстрее. Существенными являются те же компоненты, что и в первом расчете. Отличительной чертой высокой температуры является преобладание процессов диссоциации, в связи с чем нарабатывается не вода, а атомарные водород и кислород. Их концентрации монотонно возрастают.

3.3.3. Сравнение схем

Для сравнения схем вычислялась приведенная погрешность (3.13) и усреднялась по компонентам по правилу (1.13). Для нормировки полученный результат делился на сумму начальных концентраций.

На рис. 3.8 даны зависимости полученных погрешностей от числа узлов сетки для схем ERK2, ERK4 и специальной схемы (3.7) при температуре $T = 2000$ K. График дан в двойном логарифмическом масштабе.

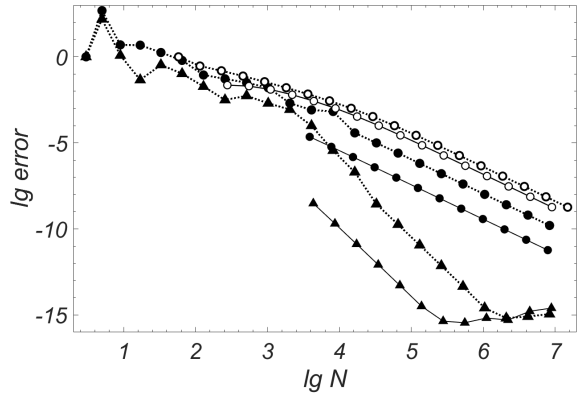

Рис. 3.8. Погрешности концентраций при сгущении сеток; $T = 2000$ K. Сплошные линии – расчеты на геометрически-адаптивных сетках, пунктирные – на равномерных сетках в длине дуги. ○ – специальная схема (3.7), ● – ERK2, ▲ – ERK4.

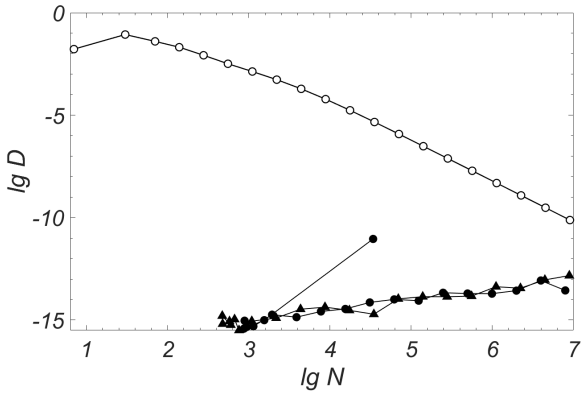

Рис. 3.9. Дисбалансы при сгущении геометрически-адаптивных сеток; $T = 2000$ K. Обозначения соответствуют рис. 3.8.

Уже при $N \approx 3000$ узлов графики погрешности выходят на прямолинейный участок, соответствующий степенному закону сходимости. Для схемы ERK2 и специальной схемы (3.7) он имеет наклон -2, а для схемы ERK4 – наклон -4. Такой наклон соответствует теоретическому порядку точности указанных схем. Это означает, что метод Ричардсона применим, и полученные оценки погрешности являются надежными. Уже при этом числе узлов специальная схема (3.7) дает точность 0.3%, схема ERK2 – 0.01%, а схема ERK4 – 0.0003%. Такой точности вполне достаточно для химических расчетов, поскольку константы реакций известны с существенно худшей точностью.

Хотя схемы ERK2 и специальная схема (3.7) имеют одинаковый порядок точности, схема ERK2 показала в ~ 100 раз лучшую фактическую точность, чем специальная схема (3.7). Строго объяснить этот результат мы не можем.

Предположительно, причиной является неконсервативность схемы (3.7), в то время как схема ERK2 строго консервативна.

Схема ERK4 оказалась еще более точной, чем схема ERK2. Ее преимущество увеличивается со сгущением сетки: на самых грубых оно составляет около 30 раз, а на подробных сетках доходит до ~ 1000 раз. Эта схема на сетках с $N \sim 10^6$ выходит на фон ошибок округления, который составляет $\sim 10^{-15}$; это лишь незначительно превышает ошибку единичного округления.

Таким образом, использование геометрически-адаптивных сеток в длине дуги позволяет применять явные схемы даже для сверхжестких задач (ярким примером которых является задача кинетики реакций). Напомним, что выигрыш по экономичности по сравнению с неявными схемами оказывается тем больше, чем больше порядок системы.

Дисбалансы. На рисунке 3.9 показаны среднеквадратичные нормы относительных дисбалансов всех трех схем при той же температуре $T = 2000$ K. Величины, отложенные на графике, вычисляются аналогично погрешностям на рисунке 3.8.

Схемы ERK2 и ERK4 консервативны, и их дисбалансы составляют $\sim 10^{-15} - 10^{-13}$. Это является следствием ошибок машинного округления (все расчеты выполнены с 64-битовыми числами). Такое поведение дисбалансов обусловлено консервативностью схем ERK2 и ERK4. При сгущении сеток дисбалансы увеличиваются из-за накопления ошибок округления. Видно, что это увеличение пропорционально $\sqrt{N}$.

Видно, что дисбаланс специальной схемы (3.7) убывает как $O(h^2)$, то есть с той же скоростью, что и погрешность. На самой грубой сетке он составляет $\sim 3\%$, что не всегда приемлемо. На подробных сетках $N \sim 10^6$ дисбаланс составляет $\sim 10^{-6}\%$, то есть ничтожно мал.

6000 K. Такое же сравнение схем проводилось при температуре $T = 6000$ K. На рисунке 3.10 приведены графики соответствующих погрешностей, на ри-

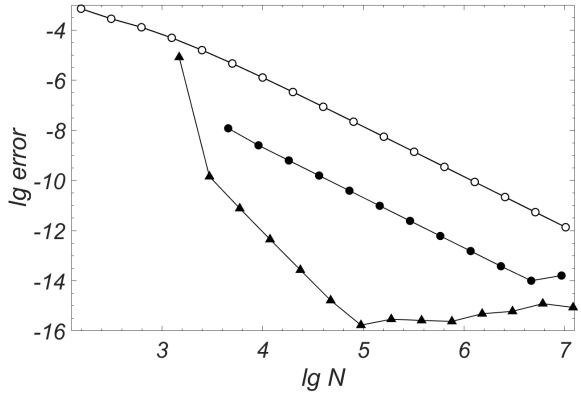

Рис. 3.10. Погрешности концентраций при сгущении геометрически-адаптивных сеток; $T = 6000$ K. Обозначения соответствуют рис. 3.8.

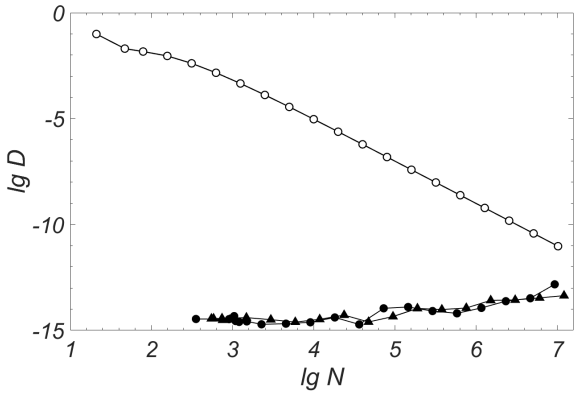

Рис. 3.11. Дисбалансы при сгущении геометрически-адаптивных сеток; $T = 6000$ K. Обозначения соответствуют рис. 3.8.

сунке 3.11 – дисбалансов. Видно, что результаты аналогичны полученным при $T = 2000$ K.

3.3.4. Равномерные сетки

Были проведены аналогичные расчеты по всем трем схемам с равномерным шагом по длине дуги. При этом на подробных сетках достаточно много узлов попадает не только регулярных участках решения, но и в пограничных слоях (включая контрастные структуры). Однако в переходных зонах число узлов оказывается невелико, и лишь на очень подробных сетках туда попадает значительное число узлов. Нормировка компонент (то есть веса в длине дуги) не учитывалась. Приведенные погрешности показаны на рис. 3.8; дисбалансы – на рис. 3.12.

Для специальной схемы (3.7) кривая погрешности состоит из двух прямолинейных участков. Начальный участок до $N \sim 20000$ имеет наклон -1. Здесь шаг еще недостаточно мал, и узлы еще не попадают в переходную зону. Поэтому фактический порядок точности понижается до первого. На втором участке $N > 20000$ наклон кривой равен -2, то есть имеет место регулярная сходимость

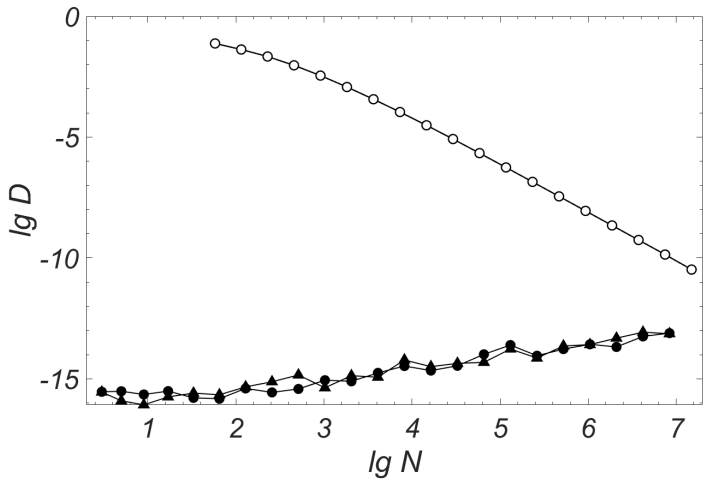

Рис. 3.12. Дисбалансы при сгущении равномерных сеток; $T = 2000$ K. Обозначения соответствуют рис. 3.8.

со вторым порядком точности. Здесь шаг настолько мал, что узлы начинают попадать внутрь переходной зоны. Это позволяет определить характерный размер последней. Он составляет $\sim 1/20000$ от полной длины дуги.

Для обеих схем Рунге-Кутты картина иная. Пока узлы не попадают внутрь переходной зоны, регулярной сходимости нет. Кривые имеют нерегулярный волнообразный характер. После того, как шаг становится достаточно малым и переходная зона начинает разрешаться, сходимость становится регулярной с теоретическим порядком точности (2-м для ERK2 и 4-м для ERK4).

Сравним количественную точность, получаемую на геометрически-адаптивной сетке и на равномерной в длине дуги. По рис. 3.8 видно, что для специальной схемы (3.7) выбор шага по кривизне уменьшает погрешность в ~ 3 раза. Для схемы ERK2 выигрыш по точности составляет ~ 30 раз, а для схемы ERK4 – 3000 раз! Это показывает высокую эффективность предложенного алгоритма выбора шага.

Заметим, что, несмотря на фактическое отсутствие сходимости на начальном участке, дисбалансы по схемам ERK2 и ERK4 остаются на уровне ошибок округления, хотя и несколько увеличиваются с ростом N. Это очевидно: для консервативных схем нельзя пользоваться контролем точности по дисба-

лансам. Необходим контроль фактической погрешности по методу Ричардсона. Дисбаланс схемы (3.7) убывает со скоростью $O\left(h^2\right)$, как и погрешность.

3.3.5. Сравнение с известными алгоритмами выбора шага

Сравним пакеты GEAR и DOPRI5 с методом геометрически-адаптивного выбора шага на задаче кинетики горения водорода в кислороде. Ранее такого исследования стандартных пакетов не проводилось, поскольку не было достаточно надежного контрольного метода, с которым можно было бы сравнить расчет по стандартному пакету.

В данном работе в качестве такого метода выбраны схемы РК на геометрически-адаптивных сетках. При этом сходимость решения контролируется по методу Ричардсона. Сгущение сеток прекращается тогда, когда погрешность достигает ошибок машинного округления, которые в данной задаче составляют $\sim 10^{-15}$ (см. рис. 3.8 и 3.10). Поэтому это решение можно считать точным. Решение по стандартным пакетам осуществляется двумя способами: в аргументе t и в аргументе l.

Процедура сравнения. Полученные по стандартным пакетам решения сравниваются с решением, полученным на геометрически-адаптивных сетках. Мерой различия служит их среднеквадратичная разность по всем узлам решения, полученного по стандартному пакету

$$D_j = \sqrt{\frac{1}{M+1}\sum_{m=0}^{M}\left(U_j - \tilde{U}_j\right)^2}, \quad D = \sqrt{\frac{1}{J}\sum_{j=1}^{J} D_j^2}. \tag{3.20}$$

Здесь U_j и $\tilde{U}_j$ – j-е компоненты решений, полученных геометрически-адаптивным методом и по стандартным пакетам соответственно, M – число шагов в решении, полученном по стандартным пакетам. Отметим, что, как правило, $M \ll N$, где N – число шагов в «точном» решении, полученном геометрически-адаптивным методом. Величины U_j и $\tilde{U}_j$ нормированы на веса ν_j, поэтому разность решений безразмерна.

Узлы геометрически-адаптивных сеток и узлы сеток, выданных стандартными пакетами, отличаются. Поэтому в (3.20) следует использовать интерполяцию U_j, в которую подставляются узлы сеток, выданных стандартными пакетами.

Были рассмотрены два способа интерполяции: кусочно-линейный и по Эрмиту с сохранением непрерывности первой производной [181]. Напомним формулы этих методов.

Кусочно-линейная интерполяция строит непрерывную кусочно-гладкую функцию, соединяя прямыми отрезками точки, известные по заданным сеточным значениям функции. На отрезке между двумя соседними точками L_{n+1} и L_n, в которых заданы сеточные значения, интерполированная функция будет иметь вид

$$\mathbf{U}(l) = \frac{\mathbf{U}_{n+1} - \mathbf{U}_n}{L_{n+1} - L_n}(l - L_n) + \mathbf{U}_n, \quad l \in [L_n, L_{n+1}]. \tag{3.21}$$

Эрмитова интерполяция сохраняет гладкость во всех точках. На такой интерполяции можно ожидать более высокой точности. Чтобы построить эрмитову интерполяцию между двумя соседними точками L_{n+1} и L_n, нужно знать в этих двух точках значения решения $\mathbf{U}$ и его первой производной $\mathbf{F}$ (то есть правой части):

$$\mathbf{U}(l) = \mathbf{a} + \mathbf{b}\xi + \left(\xi^2 - \frac{1}{4}\right)(\mathbf{c} + \mathbf{d}\xi), \; l \in [L_n, L_{n+1}], \tag{3.22}$$

$$\mathbf{a} = \frac{1}{2}(\mathbf{U}_{n+1} + \mathbf{U}_n), \quad \mathbf{b} = (\mathbf{U}_{n+1} - \mathbf{U}_n); \tag{3.23}$$

$$\mathbf{c} = \frac{1}{2}(\mathbf{F}_{n+1} - \mathbf{F}_n)(L_{n+1} - L_n), \tag{3.24}$$

$$\mathbf{d} = (\mathbf{F}_{n+1} + \mathbf{F}_n)(L_{n+1} - L_n) + 2(\mathbf{U}_{n+1} - \mathbf{U}_n), \tag{3.25}$$

$$\xi = \frac{l - \frac{1}{2}(L_{n+1} + L_n)}{L_{n+1} - L_n}. \tag{3.26}$$

Во всех случаях различие результатов расчётов с использованием кусочно-линейной и эрмитовой интерполяции было незначительным. Это обусловлено тем, что шаги в «точном» решении очень малы. Поэтому погрешность интер-

поляции в промежутках между узлами достаточно мала. Напомним, что погрешность линейной интерполяции есть $O\left(h^2\right)$, а эрмитовой – $O\left(h^4\right)$.

Обратная интерполяция длины дуги. Если решение по стандартному пакету вычислялось в аргументе «длина дуги», дополнительные действия не требуются. Если же оно вычислялось в аргументе «время», нужно предварительно найти соответствующее узлу значение длины дуги. Для этого проинтерполируем сеточную функцию $l\left(t\right)$, известную нам из «точного» решения, и вычислим её значение в узле $\tilde{t}_m$ сетки, выданной стандартным пакетом. Так как функция $t\left(l\right)$ монотонно возрастает, то она является обратимой, и производная $l\left(t\right)$ как обратной к ней функции равна

$$\frac{dl}{dt}\left(t\right) = \frac{1}{\nu_0\, F_0\left(\mathbf{U}\left(t\right)\right)}, \tag{3.27}$$

где ν_0 – вес по времени, если он используется.

Погрешность интерполяции. После нахождения «точного» решения вычисляется оценка погрешности интерполяции. Возьмём некоторый внутренний узел n и построим интерполяцию повторно по всем узлам, кроме n. Найдём значение этой интерполяции в узле n и вычислим отличие от исходного значения $\left(\mathbf{U}_j\right)_n$ в этом узле. Среднеквадратичная норма этой разности по всем внутренним узлам и берется в качестве оценки погрешности интерполяции:

$$D_j^{Int} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}\left(\left(U_j\right)_n - U_j\left(L_n\right)\right)^2}, \tag{3.28}$$

где $U_j\left(L_n\right)$ – интерполянта между узлами L_{n-1} и L_{n+1}, взятая в точке L_n; $\left(U_j\right)_n$ – сеточное значение в узле L_n. Величины U_j нормированы на веса ν_j, поэтому погрешность интерполяции безразмерна. Характерная величина погрешности интерполяции во всех расчетах составила $\sim 10^{-11}$. Это на много порядков меньше, чем различия решений. Таким образом, результатам интерполяции можно доверять, а искажением, которое она вносит в разность решений – пренебречь.

Программа GEAR. Приведём результаты вычислений при температуре $T = 6000\,\mathrm{K}$. Погрешности расчётов в аргументе t приведены на рис. 3.13. По оси

абсцисс отложена величина, обратная `RelTol`. Круглыми маркерами изображено отклонение решения по стандартному пакету от решения на геометрически-адаптивных сетках, треугольными – число узлов, получившееся в сетке стандартного пакета.

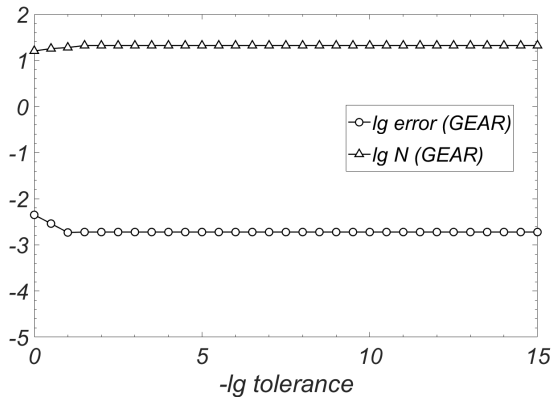

Рис. 3.13. Горение водорода в кислороде, $T = 6000$ K. Маркеры – программа GEAR в аргументе время: ∘ – фактическая погрешность, △ – число шагов сетки.

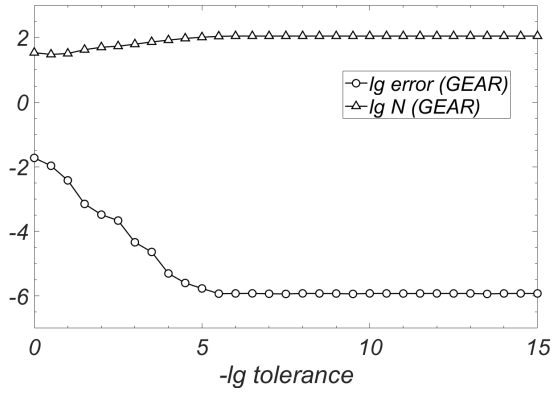

Рис. 3.14. Горение водорода в кислороде, $T = 6000$ K. Расчет по программе GEAR в аргументе длина дуги. Обозначения соответствуют рис. 3.13.

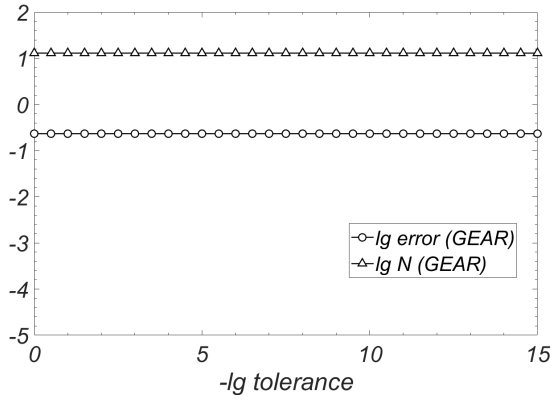

Рис. 3.15. Горение водорода в кислороде, $T = 2000$ K. Расчет по программе GEAR в аргументе время. Обозначения соответствуют рис. 3.13.

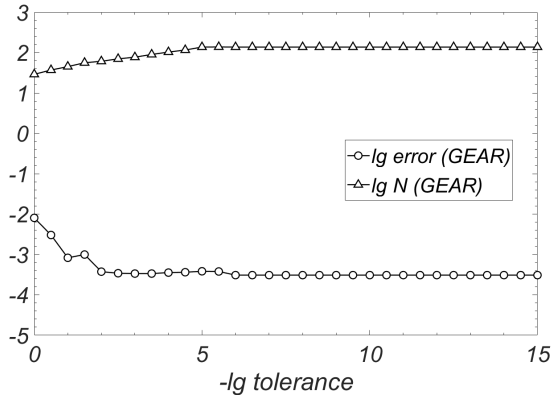

Рис. 3.16. Горение водорода в кислороде, $T = 2000$ K. Расчет по программе GEAR в аргументе длина дуги. Обозначения соответствуют рис. 3.13.

Видно, что при уменьшении `RelTol` погрешность программы Гира и выдаваемое ею число шагов быстро выходят на некоторое предельное значение и далее перестают меняться. Фактическая погрешность оказывается не лучше $5 \cdot 10^{-3}$ и совершенно не соответствует заданным значениям `RelTol`. Число узлов составляет лишь около 50, то есть трудоёмкость очень низкая. В данном примере удаётся получить точность, достаточную для практики. Однако проводить расчёты по методам, не дающим хотя бы ориентировочное представление о точности, опасно.

Рассмотрим расчёт в аргументе «длина дуги» (рис. 3.14). Качественное поведение аналогично рис. 3.13, но количественные результаты заметно лучше. Видно, что отклонение уменьшается при уменьшении `RelTol`, пока не достигнет значения 10^{-6}. При дальнейшем уменьшении `RelTol` отклонение и число узлов практически не меняются. Фактическая точность стандартных пакетов не превышает 10^{-6}, но для практических расчётов этого более чем достаточно. Число узлов, при котором достигается сходимость, составляет около 110. Это в два раза больше, чем в предыдущем случае, но всё равно немного.

Рассмотрим аналогичные результаты для расчётов по программе Гира при температуре $T = 2000\,\mathrm{K}$. На рис. 3.15 приведены результаты расчётов в аргументе t, а на рис. 3.16 – в аргументе l.

Здесь все результаты качественно похожи на приведённые выше, но количественно они хуже, так как задача становится более жёсткой. Погрешность в аргументе «время» составляет более 0.1, что может быть неудовлетворительно для практических расчётов. Точность в аргументе «длина дуги» оказывается не лучше чем $3 \cdot 10^{-4}$, но для практических расчётов этого, как правило, достаточно.

Программа Дормана-Принса. Рассмотрим теперь результаты расчётов по программе Дормана-Принса в аргументе «длина дуги». На рис. 3.17 приведены результаты расчётов при температуре $T = 6000\,\mathrm{K}$, а на рис. 3.18 – при температуре $T = 2000\,\mathrm{K}$. В аргументе «время» эта программа не сработала.

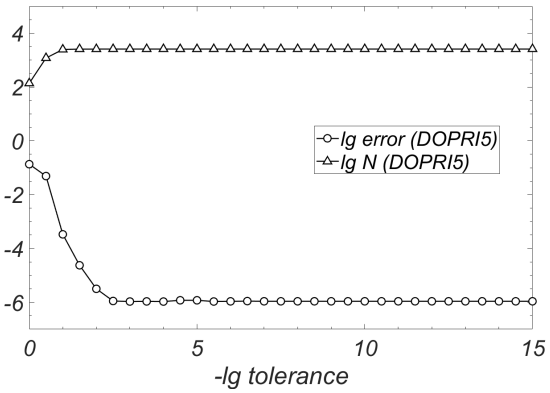

Рис. 3.17. Горение водорода в кислороде, $T = 6000$ K. Расчет по программе DOPRI5 в аргументе длина дуги. Обозначения соответствуют рис. 3.13.

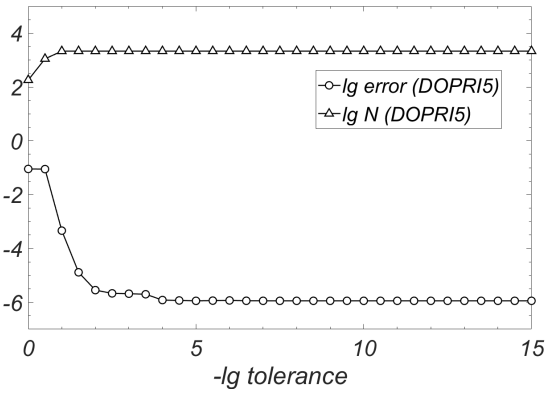

Рис. 3.18. Горение водорода в кислороде, $T = 2000$ K. Расчет по программе DOPRI5 в аргументе длина дуги. Обозначения соответствуют рис. 3.13.

Программа Дормана-Принса в целом даёт результаты, аналогичные результатам по программе Гира. Она делает подробную сетку около 3000 узлов, что существенно больше, чем в программе Гира. Однако сравнивать трудоёмкости этих программ только по числу шагов затруднительно, так как один шаг программы Гира намного дороже, чем один шаг программы Дормана-Принса. Видно, что при 6000 K и 2000 K результаты почти одинаковы. Более того, в обоих случаях фактическая точность Дормана-Принса достаточно быстро достигает 10^{-6}, чего не получилось в случае программы Гира при температуре 2000 K.

Итак, программа Дормана-Принса при использовании параметризации через длину дуги подходит для прикладных расчётов, требующих высокой точности. Программа Гира оказывается более надёжной, хотя и менее точной. Однако подчеркнём, что сделать вывод о фактической количественной точности мы смогли только после сравнения численного решения с решением, полученным другим методом с заведомо достаточной точностью.

3.4. Основные результаты главы

1. Предложен новый специализированный численный метод для задач кинетики. Этот метод явный, и его трудоемкость очень мала. Показано, что для данного типа задач этот метод является более точным и надежным, чем другие схемы первого и второго порядка точности (специализированная схема Калиткина-Гольдина и общие схемы Эйлера и Розенброка). Метод позволяет проводить вычисления одновременно с нахождением гарантированной оценки математической погрешности и пригоден к включению в газодинамические пакеты программ.
2. Проведены расчеты задачи кинетики химических реакций на примере водород-кислородного горения. Показано, что явные схемы Рунге-Кутты на геометрическ адаптивных сетках успешно справляются с этими задачами.
3. Проведено тестирование традиционных алгоритмов выбора шага на этой задаче, причем впервые контролировалась фактическая точность расчета. В качестве контрольного метода использовались явные схемы Рунге-Кутты на геометрически-адаптивных сетках. Показано, что традиционные программы теряют надежность, не обеспечивают заданную точность, а в ряде случаев и вовсе не позволяют провести расчет.

4. Задачи Коши с сингулярностями решения

4.1. Разрушение решений дифференциальных уравнений

Одной из трудных проблем, возникающих при численном расчете задачи Коши для систем ОДУ, является возникновение сингулярностей (особых точек). Обычно положение сингулярностей априори неизвестно. Эйлер заметил, что при приближении к сингулярностей уравнения Риккати приближенное решение все более и более удаляется от точного. Поэтому для исследования сингулярностей широко применяются методы компьютерной алгебры.

Вопрос о полностью численном обнаружении и исследовании сингулярностей в решении задач Коши был поставлен на семинаре Марчука в Институте вычислительной математики РАН в 2003 году.

Приведем некоторые сведения из аналитической теории дифференциальных уравнений. Они позволяют качественно исследовать поведение решения по виду правой части. Они могут быть полезны для теоретического обоснования численных методов обнаружения сингулярностей.

Рассмотрим уравнение $du/dt = F(u,t)$, причем переменные u и t будем считать комплексными. Область существования решения определяется точками, где нарушается аналитичность функции $u(t)$. Пенлеве предложил классификацию особых точек аналитических функций [4]. Приведем ее согласно [247].

Определение 2. *Особая точка $t = t_0$ функции $u(t)$ называется критической особой точкой, если при обходе этой точки значение функции $u(t)$ меняется. В противном случае особая точка t называется некритической.*

Пусть n – наименьшее целое число ($n > 1$), такое, что после n-кратного обхода точки $t = t_0$ значение функции $u(t)$ возвращается к первоначальному значению. Тогда при введении новой переменной по формуле $t = t_0 + \xi^n$ функция $u(\xi)$ является однозначной функцией переменной ξ. Тогда если вблизи $\xi = 0$ справедливо разложение

$$u(\xi) = a_0 + a_1\xi + a_2\xi^2 + \ldots, \tag{4.1}$$

то $u(t)$ выражается через t в виде

$$u(t) = a_0 + a_1(t - t_0)^{1/n} + a_2(t - t_0)^{2/n} + \dots \tag{4.2}$$

Такая точка t_0 называется критической алгебраической точкой.

Если $u(\xi)$ представима в виде

$$u(\xi) = a_{-m}\xi^{-m} + \dots + a_{-1}\xi^{-1} + a_0 + \dots, \tag{4.3}$$

то для $u(t)$ имеем

$$u(t) = a_{-m}(t - t_0)^{-m/n} + \dots + a_{-1}(t - t_0)^{-1/n} + a_0 + \dots, \tag{4.4}$$

и точка $t = t_0$ называется критическим полюсом. Такую особую точку также часто называют точкой ветвления.

Если особая точка $t = t_0$ не является критической и $u(t)$ представимо рядом (4.4), то точка $t = t_0$ является полюсом кратности m.

Определение 3. *Алгебраическими особыми точками функции $u(t)$ называются критические алгебраические особые точки, критические полюсы и простые полюсы.*

Пусть $t = t_0$ есть неалгебраическая особая точка функции $u(t)$ и пусть Δ_ρ является замыканием множества значений функции $u(t)$, которое она принимает в окрестности $\rho > 0$ точки t_0.

Определение 4. *Множество $E_{z_0} = \lim_{\rho\to 0} \Delta_\rho$ в случае монотонной зависимости Δ_ρ от ρ будем называть областью неопределенности функции $u(t)$ в особой точке $t = t_0$.*

Определение 5. *Особая точка z_0 функции $u(t)$ называется трансцендентной особой точкой, если множество неопределенности E_{z_0} состоит из одной точки.*

Определение 6. *Особая точка t_0 функции $u(t)$ называется существенно особой точкой, если множество неопределенности E_{z_0} содержит более одной точки.*

Приведем примеры. Функция $u = \sqrt{t}$ имеет критическую особую точку $t = 0$. Функция $u = 1/t$ имеет простой полюс при $t = 0$. Функция $u = \ln t$ имеет трансцендентную особую точку $t = 0$, поскольку при подходе к этой точке по любому пути u стремится к бесконечности (область неопределенности состоит из одной точки $u = \infty$). Функция $u = e^{1/t}$ имеет существенно особую точку $t = 0$.

Определение 7. *Особые точки решений дифференциальных уравнений на комплексной плоскости, положение которых зависит от начальных данных, называются подвижными особыми точками. В противном случае особые точки называются неподвижными.*

Поведение решения в окрестности подвижной особой точки было исследовано Пенлеве. Приведем формулировки теорем Пенлеве согласно [216].

Теорема 10. *Подвижные особые точки решения ОДУ первого порядка*

$$F(du/dt, u, t) = 0, \quad F \in \mathbb{Q}[v, u, t] \tag{4.5}$$

всегда является алгебраической, то есть в окрестности особой точки такое решение может быть разложено в ряд Пюизе

$$u = C(t - t_0)^p + \ldots, \quad p \in \mathbb{Q}. \tag{4.6}$$

Эта теорема допускает обобщение на многомерный случай. Рассмотрим систему ОДУ

$$\frac{du_1}{f_1} = \ldots = \frac{du_n}{f_n} = \frac{dt}{f_0}. \tag{4.7}$$

Здесь $f_0, \ldots, f_n$ – многочлены из $\mathbb{Q}[t, u_1, \ldots, u_n]$. Равенства

$$f_1(u_1, \ldots, u_n, t) = 0, \ldots, f_n(u_1, \ldots, u_n, t) = 0 \tag{4.8}$$

задают набор из $n+1$ гиперповерхности. Пересечение этих гиперповерхностей есть решение системы (4.8). Если это множество решений является конечным, то система (4.7) называется неособенной. Справедлива

Теорема 11. *Подвижная особая точка $t = t_0$ неособенной системы (4.7) является всегда алгебраической, то есть в ее окрестности решение может быть разложено в ряд Пюизе*

$$u_1 = C(t - t_0)^p + \ldots, \quad p \in \mathbb{Q}. \tag{4.9}$$

Обзор наиболее известных методов расчета задач Коши с сингулярностями решения приведен в приложении.

4.2. Обнаружение ближайшей сингулярности

4.2.1. Алгебраическая особая точка

Скалярное уравнение. Поясним основную идею на примере одного уравнения (задача Марчука)

$$du/dt = f(u), \quad f(u) = u^\nu, \quad \nu > 1, \quad u(0) = u_0. \tag{4.10}$$

Далее будет показано, что этот метод без труда обобщается на системы уравнений. Строгое обоснование предлагаемого подхода будет дано в п. 4.4. Точное решение задачи (4.10)

$$u(t) = u_0(1 - t/t_0)^{-1/(\nu-1)}. \tag{4.11}$$

имеет алгебраическую особую точку порядка $q = (\nu-1)^{-1}$ в точке $t_0 = u_0^{1-\nu}/(\nu-1)$. Формально решение по любой явной либо явно-неявной схеме не имеет вертикальной асимптоты, однако фактически оно очень быстро нарастает вблизи особенности. На рис. 4.1 представлено решение задачи (4.10) по схеме одностадийной схеме Розенброка с комплексным коэффициентом (CROS) для $q = 1/2$

($\nu = 3$) на разных сетках по l. После прохождения точки t_0 решение круто уходит вверх. Чем мельче шаг сетки, тем ближе оно ложится к асимптоте точного решения.

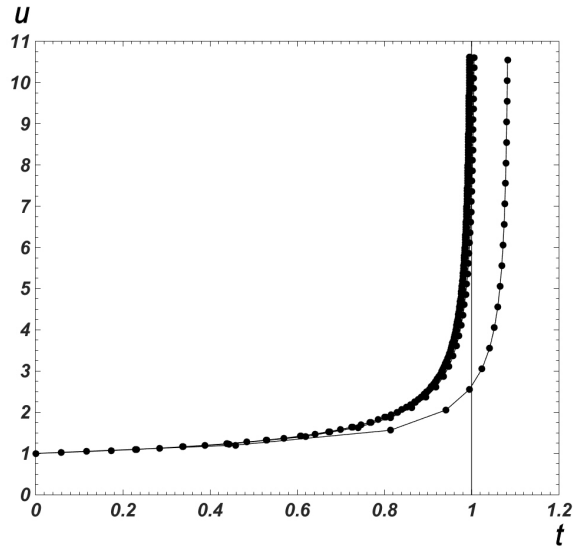

Рис. 4.1. Решения задачи (4.10) для $q = 1/2$, $t_0 = 1$ на сгущающихся сетках. Маркеры – расчетные точки, вертикальная линия – асимптота точного решения (4.11).

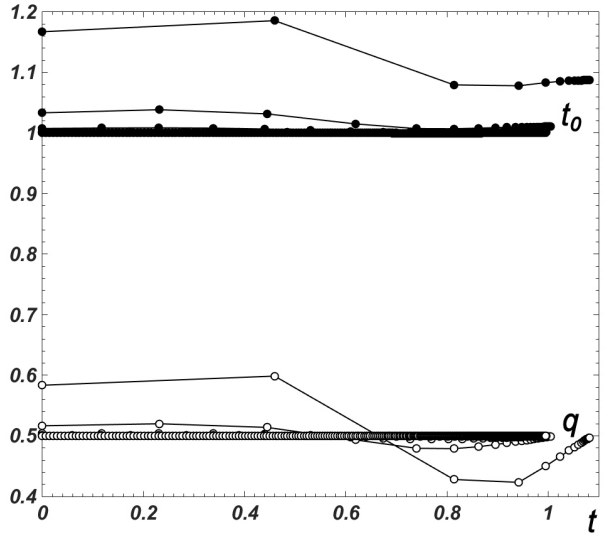

Рис. 4.2. Профили q и t_0 на сгущающихся сетках в задаче (4.10).

Согласно теореме 10, в окрестности точки t_0 точное решение с алгебраической особой точкой представимо рядом Пюизе

$$u(t) = C(t_0 - t)^{-q} + \dots \tag{4.12}$$

Дифференцируя его, получим

$$f = qu/(t_0 - t). \tag{4.13}$$

Это соотношение справедливо при любых аргументах, и в частности, в расчетные моменты времени t_n. Записав его в узлах n и $n + 1$, получим систему алгебраических уравнений относительно q и t_0. Ее решение имеет вид

$$q = \frac{t_{n+1} - t_n}{u_n/f_n - u_{n+1}/f_{n+1}}, \quad t_0 = q\frac{u_n}{f_n} + t_n. \tag{4.14}$$

Выражения (4.14) не зависят от параметризации интегральных кривых, поэтому они применимы как при аргументе t, так и при аргументе l.

Будем вычислять значения q и t_0 во всех узлах сетки, на рис. 4.2 показаны профили этих величин на сгущающихся сетках. Можно визуально наблюдать, как они сходятся к теоретическим значениям при увеличении n. В прикладных вычислениях рекомендуется строить аналогичные графики.

В аргументе t расчеты по явным и явно-неявным схемам можно формально вести до $t = +\infty$ (поскольку на каждом n-м шаге можно вычислить u_{n+1}). Однако довольно быстро численное решение выходит за пределы представимых чисел, и происходит переполнение. Создается впечатление, что особая точка еще не достигнута, хотя на самом деле может быть уже давно пройдена. При расчетах по чисто неявным схемам численное решение за особенностью может не существовать (так как нелинейное уравнение относительно u_{n+1} может не иметь решений). Все это делает расчеты в аргументе t неудобными и ненадежными. Намного удобнее использовать длину дуги, которая неограниченно растет вдоль интегральной кривой. На преимущества длины дуги в этой проблеме указал Марчук в 2005 году.

Если на достаточно подробных сетках с увеличением текущей длины дуги значения q и t_0, определяемые из (4.14), выходят на константы, то поведение решения определяется множителем $(t_0 - t)^{-q}$, и можно надежно диагностировать алгебраическую особую точку.

В выражения (4.14) входят только сеточные значения u (и f, которые выражаются через них). Поэтому погрешности вычисления q и t_0 определяются только погрешностью u. Таким образом, к q и t_0 можно применить стандартные оценки погрешности по методу Ричардсона (см. п. 1.4). Например, для q имеем

$$\Delta q = \frac{q_N - q_{rN}}{r^p - 1}, \tag{4.15}$$

где N – число узлов более грубой сетки, r – кратность сгущения сетки, p – порядок точности используемой схемы. Оценка для t_0 аналогична. Поскольку при

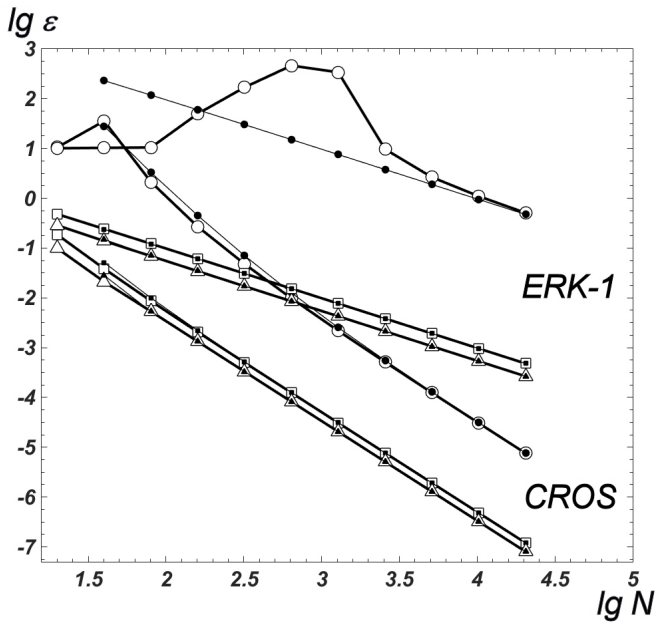

Рис. 4.3. Сходимость в задаче (4.10); $\circ$ – u, $\triangle$ – q, $\square$ – t_0; светлые маркеры – погрешность по точному решению; темные маркеры – оценки точности по методу Ричардсона. Названия схем указаны у кривых.

применении этой процедуры сеточные функции обычно сравнивают поточечно (т.е. в совпадающих узлах сеток N и rN), то практически всегда берут $r = 2$.

На рис. 4.3 представлены кривые сходимости u, q, t_0 при сгущении сеток по длине дуги для комплексной схемы Розенброка CROS и явной схемы Рунге-Кутты первого порядка (ERK1) в зависимости от N. График дан в двойном логарифмическом масштабе, так что степенному характеру зависимости погрешности от числа N шагов сетки соответствует прямая линия. Наклон этой прямой равен порядку точности схемы.

Видно, что расчетные значения q и t_0 сходятся к теоретическим, причем порядок сходимости равен порядку точности схемы (второй для схемы CROS и первый для ERK1). Оценки погрешности q и t_0 по методу Ричардсона отлично совпадают с погрешностями на точных значениях этих величин (даже для схемы ERK1, у которой точность самого решения очень низкая). Поэтому данная методика исключительно надежна.

Практические рекомендации. Для проведения диагностики от схемы требуется одновременно хорошая точность и высокая надежность. Однако явные

схемы Рунге-Кутты высокого порядка точности требуют слишком малого шага. Явная схема первого порядка – так как ее точность невелика, а схемы более высокого порядка – из-за невысокой надежности. Поэтому расчеты по явным схемам трудоемки.

Явно-неявная схема CROS4 также оказалась недостаточно надежной несмотря на свои формально высокие теоретические показатели (точность $O(h^4)$ и L_4-устойчивость). Скорее всего, это связано с тем, что каждая из стадий этой схемы не является даже A-устойчивой.

Только схема CROS сочетала неплохую точность $O(h^2)$, L_2-устойчивость и очень высокую надежность. Эта совокупность свойств позволяет рекомендовать именно данную схему для задач диагностики.

Система уравнений. В случае системы ОДУ значения q и t_0 следует вычислять для каждой компоненты, при этом из всех моментов времени $t_0^{(j)}$ следует выбрать наименьший. Например, рассчитывалась система

$$\frac{du}{dt} = \frac{a}{v}, \quad \frac{dv}{dt} = -\frac{b}{u}, \quad u(0) = u_0, \quad v(0) = v_0. \tag{4.16}$$

В ней v имеет алгебраическую особую точку порядка $q^{(v)} = b/(a-b)$, а u – нуль порядка $q^{(u)} = -a/(a-b)$ в одной и той же точке $t_0 = -u_0 v_0/(a-b)$. Выберем $a_0 = -1.5$, $b_0 = -0.5$, $u_0 = v_0 = 1$; тогда $t_0 = 1$, $q^{(v)} = 0.5$, $q^{(u)} = 1.5$.

Система (4.16) рассчитывалась по схеме CROS в длине дуги. На рис. 4.4 показаны профили t_0, $q^{(u)}$, $q^{(v)}$ на сгущающихся сетках, вычисленные по формулам (4.14). Видно, что поведение профилей t_0 и $q^{(v)}$ аналогично описанному в предыдущем примере, а профили $q^{(u)}$ сходятся к постоянному значению, равному -1.5. Это объясняется тем, что нуль можно рассматривать как алгебраическую особую точку отрицательного порядка. Таким образом, поведение численного решения для u и v соответствует теоретическому.

Исследовалось поведение погрешностей в норме C величин t_0, $q^{(u)}$, $q^{(v)}$ на точном решении и их оценок по методу Ричардсона в зависимости от числа узлов N. Оно аналогично расчету по схеме CROS в предыдущем примере. При

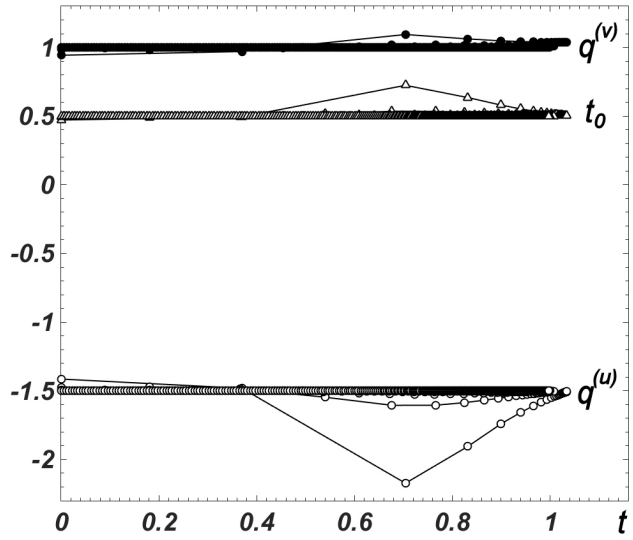

Рис. 4.4. Профили $q^{(u)}$, $q^{(v)}$, t_0 на сгущающихся сетках в задаче (4.16).

сгущении сетки по l имела место сходимость с порядком точности $p = 2$; оценки по Ричардсону отлично согласовались с точными значениями погрешности как для u, так и для t_0, $q^{(u)}$, $q^{(v)}$.

Таким образом, предлагаемая процедура позволяет диагностировать алгебраическую особую точку и для систем уравнений и вычислять ее характеристики с гарантированной точностью. Данная процедура переносится и на другие типы особенностей, описанные ниже.

4.2.2. Логарифмическая особенность

Метод диагностики. Пусть решение вблизи полюса t_0 точное решение имеет вид

$$u \approx C[\ln(t_0 - t)]^q, \quad C = \text{const}. \tag{4.17}$$

Такая особая точка является трансцендентной. Дифференцируя (4.17), нетрудно получить выражение, связывающее f и u:

$$f = -\frac{qu}{(t_0 - t)\ln(t_0 - t)}. \tag{4.18}$$

Записывая (4.18) в узлах n и $n+1$, получим систему алгебраических уравнений относительно t_0 и q. Она преобразуется к виду

$$\frac{f_n}{u_n}(t_0 - t_n)\ln(t_0 - t_n) = \frac{f_{n+1}}{u_{n+1}}(t_0 - t_{n+1})\ln(t_0 - t_{n+1}), \tag{4.19}$$

$$q = -\frac{f_n}{u_n}(t_0 - t_n)\ln(t_0 - t_n). \tag{4.20}$$

Лемма 1. *Уравнение (4.19) имеет не менее двух вещественных корней относительно* t_0. •

Доказательство. Приведем (4.19) к более удобному виду

$$0 = \frac{1}{a}x\ln x - (x+\tau)\ln(x+\tau) \equiv \varphi_1(x) - \varphi_2(x). \tag{4.21}$$

Здесь $a = (f_n u_{n+1})/(f_{n+1}u_n)$, $x = t_0 - t_n > 0$, а $\tau = t_{n+1} - t_n$ – шаг по времени. Вблизи особенности $|f|$ нарастает быстрее, чем $|u|$. Кроме того, все f и u имеют один и тот же знак, а так же $|f_{n+1}| > |f_n|$, $|u_{n+1}| > |u_n|$. Поэтому $0 < a < 1$.

Рассмотрим область малых $x \to 0$, то есть текущая точка t_n находится очень близко от t_0. Тогда $\varphi_1(x) \to 0 - 0$. При этом $\varphi_2(x) \to \tau\ln\tau < 0$ при фиксированном $\tau < 1$. Поэтому $\varphi_1(x) > \varphi_2(x)$ в некоторой окрестности $x = 0$.

Теперь рассмотрим область "средних" x, лежащих вне упомянутой окрестности $x = 0$, но по-прежнему много меньших τ. Легко видеть, что при фиксированных a и τ существует область x, в которой $x\ln x < a\tau\ln\tau$. При таких x $\varphi_1(x) < \varphi_2(x)$. Поэтому на границе областей "малых" и "средних" x имеется корень $x_{\text{м}} < \tau$.

Наконец, рассмотрим область "больших" $x \gg \tau$. Здесь $\varphi_2(x) \approx x\ln x$. Поэтому с учетом того, что $a < 1$, немедленно получаем $\varphi_1(x) > \varphi_2(x)$. Это означает, что имеется еще один корень $x_{\text{б}} > \tau$. ■

Таким образом, доказано существование не менее двух корней уравнения (4.21). Из них следует выбирать тот, который обеспечивает выход q и t_0 на стационары по мере приближения к особенности. На практике проще всего находить их методом Ньютона. Но тогда встает вопрос о выборе начального приближения. В начальный момент времени значение $x = t_0 - t_n$ велико (порядка

полного времени расчета T либо полной длины дуги L). Целесообразно выбрать эти величины в качестве начального приближения, и тогда метод Ньютона сойдется к большему корню $x_б$.

В последующие моменты времени хорошим начальным приближением будет значение x, полученное в предыдущий момент времени. Оно попадает в τ-окрестность искомого корня, и на достаточно подробных сетках метод Ньютона демонстрирует быструю сходимость.

Пример. Рассмотрим задачу

$$\frac{du}{dt} = qu^{1-1/q} \exp\{u^{1/q}\}. \tag{4.22}$$

Точное решение имеет вид $u = [-\ln(t_0 - t)]^q$. При $t < t_0 < 1$ логарифм оказывается отрицательным, и в степень возводится положительное число. Поэтому все вычисления оказываются чисто вещественными. Момент особенности t_0 определяется начальными условиями. Мы брали $q = 1.5$, $t_0 = 0.5$ и подгоняли начальные условия под эти значения.

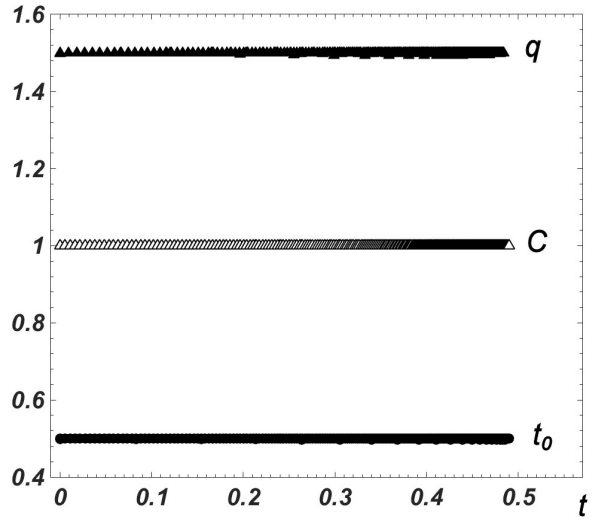

Рис. 4.5. Профили q, t_0 и C на сгущающихся сетках в задачах (4.22) и (4.23).

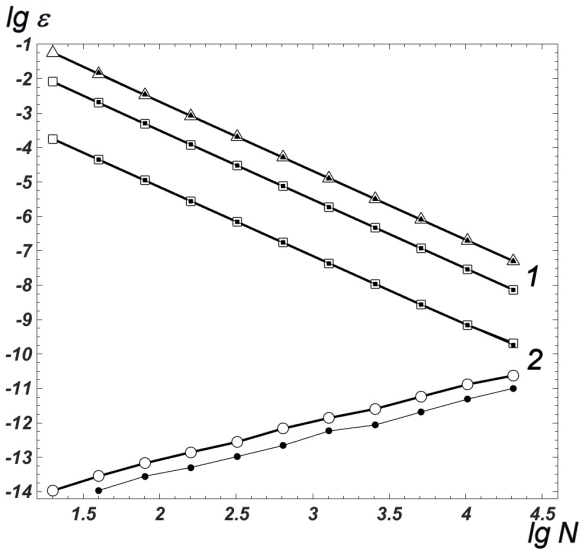

Рис. 4.6. Сходимость: 1 – в задаче (4.22), △ – q, □ – t_0; 2 – в задаче (4.23), ○ – C, □ – t_0; светлые маркеры – погрешность по точному решению, темные – оценки по точному решению.

Помимо указанного решения имеется также тривиальное $u = \text{const}$, пересекающееся с ним в начальный момент времени. При расчетах по не слишком надежным схемам численное решение может "садиться" на это тривиальное решение. Мы пользовались высоконадежной схемой CROS в длине дуги, в которой таких трудностей не возникало.

Результаты расчетов представлены на рис. 4.5, 4.6. Расчетные профили t_0 и q на сгущающихся сетках выходят на постоянные, равные соответствующим теоретическим значениям. Это доказывает, что характер особенности логарифмический. Погрешности в норме C найденных значений t_0 и q в зависимости от числа узлов N в двойном логарифмическом масштабе выходят на прямые линии с наклоном 2. Поэтому порядок сходимости равнялся $p = 2$, что соответствует теоретическому порядку сходимости для схемы CROS. Кроме того, погрешности по методу Ричардсона хорошо совпадают с погрешностями, определенными непосредственным сравнением с известным точным решением.

Частный случай. При $q = 1$ имеем $u \approx C \ln(t_0 - t)$. Тогда соотношение (4.18) существенно упрощается

$$f = -C \exp\{-u/C\}. \tag{4.23}$$

Записывая (4.23) в узлах n и $n+1$ и деля одно выражение на другое, получим простое уравнение относительно C

$$\frac{f_n}{f_{n+1}} = \exp\{(u_{n+1} - u_n)/C\}, \tag{4.24}$$

откуда

$$C = \frac{u_{n+1} - u_n}{\ln f_n - \ln f_{n+1}}. \tag{4.25}$$

Далее, пользуясь видом $u(t)$ и полученным значением C, находим t_0

$$t_0 = \frac{t_n - t_{n+1} f_{n+1}/f_n}{1 - f_{n+1}/f_n}. \tag{4.26}$$

Данный частный случай интересен постольку, поскольку для C и t_0 удается построить явные выражения, проверка которых удобна в практических расчетах.

Если профили C и t_0, вычисленные по формулам (4.25) – (4.26), выходят на постоянные значения, то можно утверждать, что в момент времени t_0 имеет место логарифмический полюс первого порядка. Сама же величина t_0 определяется начальными условиями. В демонстрационном расчете мы брали $C = 1$, $t_0 = 0.5$.

Как и в предыдущих расчетах, здесь использовалась схема CROS в длине дуги. Профили C и t_0 на сгущающихся сетках показаны на рис. 4.5. Видно, что они отлично ложатся на постоянные значения, равные 1 и 0.5 соответственно. При этом значение t_0 сходится к теоретическому со вторым порядком точности, так как кривая погрешности в двойном логарифмическом масштабе представляет из себя прямую линию с наклоном 2 (см. рис. 4.6). Погрешности величины C принимают очень малые значения (порядка $10^{-14} \div 10^{-10.5}$) и нарастают при увеличении числа узлов N. Причина этого в том, что величина C совпадает со своим теоретическим значением с точностью до ошибок округления, которые нарастают при сгущении сеток.

4.2.3. Смешанная особенность

Метод диагностики. Данная особенность представима в виде

$$u \approx C\frac{\ln(t_0 - t)}{(t - t_0)^q}, \quad C = \text{const}. \tag{4.27}$$

В этом случае имеем

$$f = -\frac{u}{(t_0 - t)\ln(t_0 - t)} + \frac{qu}{t_0 - t}. \tag{4.28}$$

Записывая (4.28) в узлах n и $n + 1$, получаем алгебраические уравнения относительно t_0 и q

$$\begin{gathered}[(t_0 - t_{n+1})\ln(t_0 - t_{n+1})f_{n+1}/u_{n+1} + 1]\ln(t_0 - t_n) = \\ = [(t_0 - t_n)\ln(t_0 - t_n)f_n/u_n + 1]\ln(t_0 - t_{n+1}),\end{gathered} \tag{4.29}$$

$$q = (t_0 - t_n)\frac{f_n}{u_n} + \frac{1}{\ln(t_0 - t_n)}. \tag{4.30}$$

Лемма 2. *Уравнение (4.29) имеет не менее двух вещественных корней относительно* t_0. •

Доказательство. Перепишем (4.29) в более удобном виде

$$0 = \left[\frac{f_{n+1}}{u_{n+1}} x \ln x + 1\right] \ln(x+\tau) - \left[\frac{f_n}{u_n}(x+\tau)\ln(x+\tau) + 1\right] \ln x \equiv \varphi_1(x) - \varphi_2(x). \tag{4.31}$$

Здесь по-прежнему $x = t_0 - t_n > 0$, а $\tau = t_{n+1} - t_n$.

Пусть $x \to 0$. Тогда $\varphi_1(x) \to \ln\tau < 0$ – фиксированная величина. При этом $\varphi_2(x) \sim A \ln x$, где $A = (f_n/u_n)\,\tau \ln\tau + 1 = \text{const}$. Если сетка достаточно подробная (то есть $\tau \sim 1/N$ достаточно мало), то $A > 0$. Это значит, что $\varphi_2(x) \to -\infty$. Таким образом, в некоторой малой окрестности $x = 0$ имеем $\varphi_1(x) > \varphi_2(x)$.

Далее, положим $x = \tau$ и исследуем знак выражения (4.31). Оно преобразуется к виду

$$\varphi_1(x) - \varphi_2(x) = \left(\frac{f_{n+1}}{u_{n+1}} - 2\frac{f_n}{u_n}\right) x \ln x \ln 2x + \ln 2. \tag{4.32}$$

Если сетка достаточно подробная, то с хорошей точностью справедливы равенства $f_{n+1} = f(t_n + \tau) \approx f_n + \tau f_n'$, $u_{n+1} = u(t_n + \tau) \approx u_n + \tau f_n$. Подставляя эти выражения в (4.32), нетрудно убедиться, что выражение в круглой скобке отрицательно и равно по порядку величины $\sim -f_n/u_n + O(\tau)$. Поэтому достаточно близко от особенности первое слагаемое в (4.32) оказывается больше по модулю, чем $\ln 2 \approx 0.7$. Таким образом, в некоторой окрестности $x = \tau$ имеем $\varphi_1(x) < \varphi_2(x)$. Следовательно, существует первый (меньший) корень $x_{\text{м}} < \tau$.

Пусть теперь $x \gg \tau$ достаточно велико. Тогда $\varphi_1(x) \approx [(f_{n+1}/u_{n+1})x \ln x + 1] \ln x$, $\varphi_2(x) \approx [(f_n/u_n)x \ln x + 1] \ln x$. Выше при исследовании логарифмической особенности было замечено, что $f_{n+1}/u_{n+1} > f_n/u_n$. Поэтому $\varphi_1(x) > \varphi_2(x)$ при достаточно больших x. Отсюда следует, что существует второй (больший) корень $x_{\text{б}} > \tau$. ■

На практике корни уравнения (4.31) удобно находить методом Ньютона. Начальное приближение можно выбирать так же, как в задаче с логарифмическим полюсом (см. п. 4.2.2).

Пример. Построить автономную задачу с рассматриваемой особенностью не удалось, поэтому рассмотрим следующую неавтономную задачу:

$$\frac{du}{dt} = \left[-\frac{u}{\ln(t_0 - t)}\right]^{1+1/q} + q\frac{u}{t_0 - t}. \tag{4.33}$$

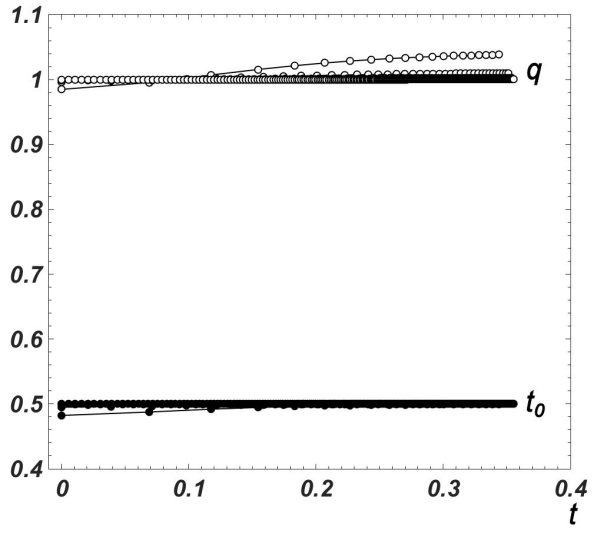

Рис. 4.7. Профили q и t_0 и C на сгущающихся сетках в задаче (4.33).

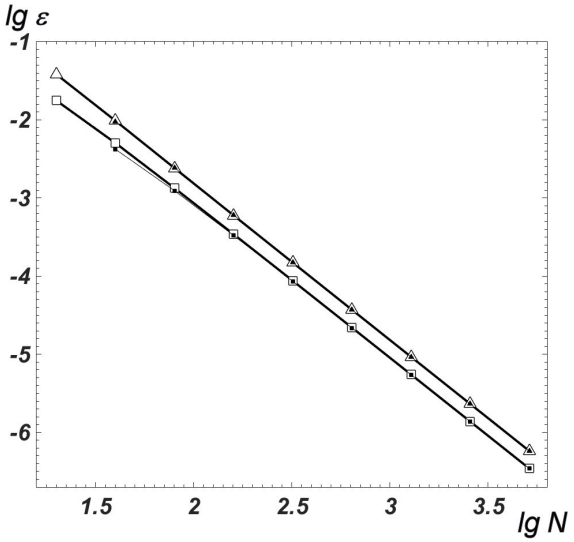

Рис. 4.8. Сходимость в задаче (4.33); обозначения соответствуют рис. 4.3.

Точное решение имеет вид $u = -\ln(t_0 - t)/(t_0 - t)^q$. При $t < t_0 < 1$ логарифм отрицателен, решение $u > 0$ положительно, поэтому все вычисления чисто вещественные. Положение особенности t_0 определяется начальными условиями. В демонстрационном расчете мы брали $q = 1$, $t_0 = 0.5$ и соответственно подгоняли начальные условия.

Как и ранее, мы использовали схему CROS в длине дуги. Результаты расчетов аналогичны предыдущим случаям (см. рис. 4.7, 4.8). Профили расчетных t_0 и q на сгущающихся сетках стремятся к постоянным значениям при увеличении n. С графической точностью эти значения совпадают с теоретическими на всех сетках, начиная со второй. Анализ дальнейших знаков показывает, что имеет

место сходимость этих величин к теоретическим при сгущении сеток, причем порядок этой сходимости равен $p = 2$.

4.2.4. Неизвестная особенность

В предыдущих разделах были подробно разобраны важнейшие виды особенностей. Однако при расчете реальных задач тип особенности заранее неизвестен. Кроме того особенность может оказаться более сложной, чем рассмотренные.

На практике рекомендуется проверять все 3 типа особенностей: вычислять значения q и t_0 по формулам (4.14), (4.19) – (4.20), (4.29) – (4.30) и строить профили этих величин. Если при выбранной гипотезе о типе особенности профили выходят на постоянные значения, то гипотеза о типе особенности подтверждается. Если же мгновенное значение q и t_0 меняется по мере приближения к особенности, то это значит, что выбранная гипотеза о ее типе неверна.

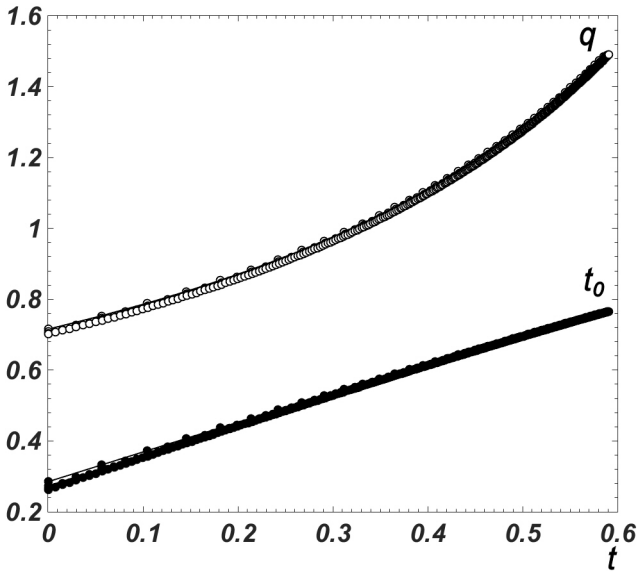

Рис. 4.9. Диагностика задачи (4.10) при $q = 2$, $t_0 = 1$ по формулам (4.19) – (4.20).

Пример такой “ошибочной” диагностики представлен на рис. 4.9. Мы взяли тестовую задачу (4.10) со степенным полюсом порядка $q = 2$ в момент $t_0 = 1$ и применили к ней формулы (4.19) – (4.20) для логарифмического полюса. В результате в расчетном промежутке времени мгновенное значение t_0 меняет-

ся в 3 раза, а мгновенное значение q – более, чем в 2 раза, причем по мере приближения к особенности оно меняется все более круто.

Если профили величин q и t_0 не выходят на строго постоянные, но их вариация вблизи особенности не слишком велика (например, не превышает 10-15%), то тип особенности близок к одному из указанных, но сама особенность отличается от него дополнительными меньшими членами. Тогда можно говорить об эффективных значениях q и t_0. В качестве них разумно брать значения в последнем узле $(q)_N$, $(t_0)_N$, который наиболее близок к особенности.

4.2.5. S-режим нелинейного горения

Нелинейное горение. Предлагаемая методика применялась к исследованию S-режима нелинейного горения, который описывается нелинейным параболическим уравнением

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^2 \frac{\partial u}{\partial x}\right) + u^3. \tag{4.34}$$

Это уравнение имеет точное решение

$$u(x,t) = \frac{\sqrt{3}}{2} \frac{1}{\sqrt{t_0 - t}} \cos\left(\frac{x}{\sqrt{3}}\right). \tag{4.35}$$

Температурные профили этого решения в фиксированные моменты времени показаны на рис. 4.10.

Поскольку переменные разделяются, то особенность типа полюс порядка $q = 1/2$ имеет место в каждой точке пространства, причем на всем отрезке решение разрушается одновременно. Иными словами, при каждом x зависимость температуры от времени имеет вид, показанный на рис. 4.1.

Методом прямых уравнение (4.34) сводится к системе ОДУ

$$\frac{du_j}{dt} = \frac{1}{2h_x^2}\left[(u_{j+1}^2 + u_j^2)(u_{j+1} - u_j) - (u_j^2 + u_{j-1}^2)(u_j - u_{j-1})\right] + u_j^3, \tag{4.36}$$

где j и h_x – номер узла и шаг по пространству соответственно. Нетрудно убедиться, что пространственный оператор в (4.34) аппроксимируется с точностью

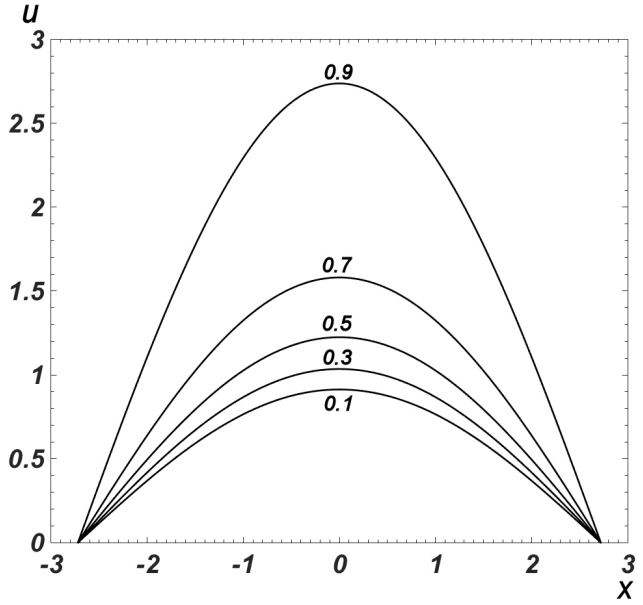

Рис. 4.10. Профили решения (4.35) в фиксированные моменты времени (указаны у кривых).

$O(h_x^2)$. Система (4.36) содержит несколько сотен компонент (для получения хорошей точности по пространству), а ее правые части весьма сложны. Поэтому такой тест достаточно представителен.

Сгущение сеток по l. Вычисления проводились по схеме CROS, имеющей второй порядок точности и обладающей L_2-устойчивостью и чрезвычайно высокой надежностью. Сетка по x содержала $J = 200$ узлов, так что система (4.36) имела огромный порядок.

На рис. 4.11 показаны кривые сходимости u, t_0, q при сгущении сеток по длине дуги. Четко видно, что все 3 величины сходятся со вторым порядком точности. Однако, начиная с некоторой достаточно подробной сетки, кривые погрешности, вычисленные сравнением с точным решением (4.35), выходят на горизонтальный участок на уровне $\sim 10^{-3}$. При этом оценки точности по Ричардсону, относящиеся к системе ОДУ (4.36), продолжают сходиться со вторым порядком вплоть до фона ошибок округления (показан для u).

Это связано с тем, что при введении дискретизации по пространству мы вносим некоторую погрешность, и поэтому полюс точного решения (4.35) отличается от полюсов системы (4.36), либо она их может вовсе не иметь. В по-

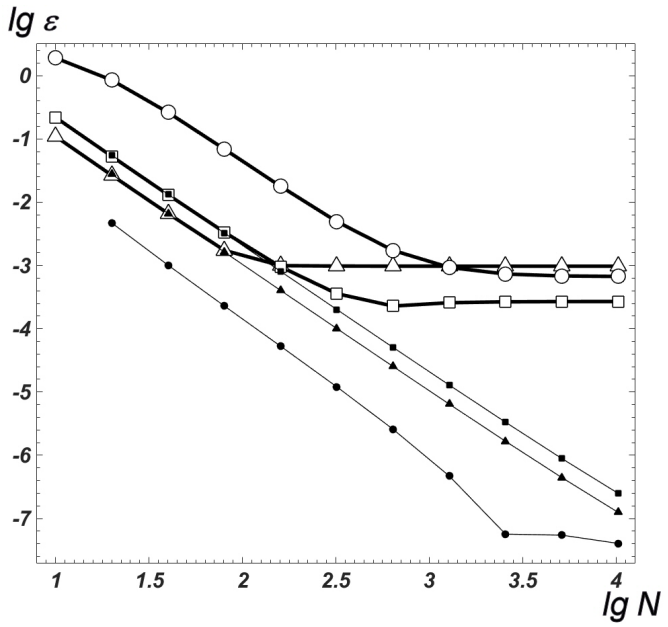

Рис. 4.11. Сходимость в задаче (4.36), расчет по схеме CROS. Обозначения соответствуют рис. 4.3. В качестве точного решения выбрано (4.35).

следнем случае профили q и t_0 – некоторые кривые не обязательно выходящие на постоянные. Чтобы уменьшить это отличие, нужно вводить более подробные сетки по x. По этой же причине погрешность на точном решении для u не сопадает с погрешностью, полученной методом Ричардсона применительно к системе (4.36). Погрешность, вносимую дискретизацией по x, можно оценить методом сгущения сеток.

Сгущение сеток по x**.** При сгущении сеток по пространству увеличивается число компонент в системе (4.36). Если вести расчет этой системы до фиксированной длины дуги, то при увеличении числа компонент длина каждой интегральной кривой уменьшится, и расчетный интервал времени сократится.

Поэтому целесообразно вести расчет до достижения заданного расчетного времени. Тогда добавление новых компонент приводит к увеличению суммарной длины всех интегральных кривых. В результате узлы сеток по l, относящиеся к разным сеткам по x, не будут совпадать. Это не позволяет проводить поточечные сравнения сеточных функций, однако мы будем сравнивать только значения $(q)_N$ и $(t_0)_N$ в N-м узле.

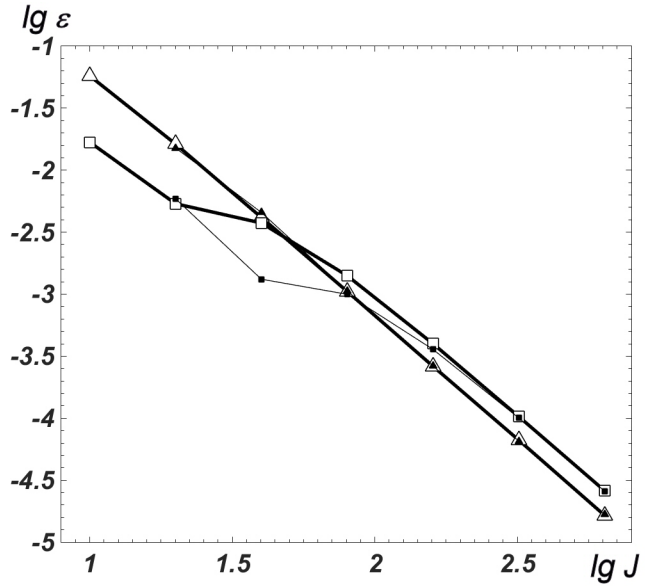

Рис. 4.12. Сходимость в задаче (4.34) при сгущении сеток по x. Обозначения соответствуют рис. 4.3.

Подчеркнем, что для исследования порядка точности по пространству сетки по l должны быть настолько подробными, чтобы погрешности q и t_0 на точном решении достигали горизонтального участка (см. рис. 4.11). Аппроксимация оператора по пространству есть $O(h_x^2)$, поэтому при сгущении сетки по x вдвое высота этого горизонтального участка должна уменьшиться в 4 раза.

На рис. 4.12 приведены кривые сходимости q и t_0 при сгущении сеток по x. Представлены как погрешности, вычисленные сравнением с точным решением, так и их оценки по методу Ричардсона. Видно, что на грубых сетках сходимость является нерегулярной, однако начиная с $J \approx 120$ кривые выходят на прямолинейные участки с наклоном $p = 2$. Это соответствует регулярной сходимости. При этом на регулярном участке оценки по методу Ричардсона отлично совпадают с фактической погрешностью, вычисленной сравнением с точным решением.

Таким образом, предложенная методика является более простой, надежной, чем известные ранее методы.

4.2.6. Пакет программ

Предлагаемый алгоритм реализован в виде прикладного пакета программ `SiDiaG` в среде Matlab. В нем реализована автоматическая диагностика ближайшей сингулярности. Программа находит численное решение и параметры его сингулярности с апостериорной асимптотически точной оценки погрешности. В пакет входят две подпрограммы для диагностики степенного и логарифмического полюсов, а также подпрограмма оценка погрешности как решения, так и характеристик сингулярности. Ранее подобное математическое обеспечение не предлагалось. Пакет `SiDiaG` распространяется по свободной лицензии BSD-3-clause и доступен по ссылке `https://github.com/ABelov91/SiDiaG`.

4.3. Множественные полюсы первого порядка

4.3.1. Продолжение за полюс

Решения задач Коши с сингулярностями выходят за рамки известных теорем о разрешимости. Сначала определим, что понимается под продолжением решения за полюс. Для наглядности рассмотрим одно автономное ОДУ

$$du/dt = f(u),\ t > 0; \quad u(0) = u_0. \tag{4.37}$$

Уравнение (4.37) интегрируется вдоль вещественной прямой. Случай систем рассматривается аналогично. Неавтономные задачи сводятся к системе ОДУ с помощью процедур автономизации.

Пусть решение (4.37) имеет произвольное число изолированных особых точек на вещественной оси. Будем считать, что $u(t)$ аналитична в некоторых проколотых окрестностях этих точек, а сами эти точки являются алгебраическими особыми точками первого порядка. Вслед за [248, 249] мы будем называть их полюсами первого порядка. В остальной части комплексной плоскости t будем

предполагать, что решение имеет столько непрерывных производных, сколько требуется для конкретной схемы интегрирования ОДУ.

Положение полюсов не известно априори. Обозначим точку ближайшего полюса через $t = t_*$.

Перейдем к комплексным значениям аргумента t. Тогда полюс на вещественной оси можно обойти по контуру на комплексной плоскости и, вернувшись на вещественную ось, продолжить решение. При этом не возникает никаких особенностей. Такой подход позволяет продолжить решение через любое число полюсов. Однако такой подход неконструктивен, так как требует существенно более трудоемкого численного интегрирования в комплексной плоскости.

Правее точки $t = t_*$ формально существует бесконечно много точных решений дифференциального уравнения (4.37). Какое из них соответствует начальному условию u_0? Наличие полюса первого порядка в точке t_* означает, что точное решение разлагается в этой точке в следующий ряд Пюизе:

$$u = \frac{a_{-1}}{t - t_*} + \sum_{p=0}^{P+1} a_p (t - t_*)^p + o\left((t - t_*)^{P+1}\right). \tag{4.38}$$

Здесь предполагается, что $f(u)$ имеет P непрерывных производных. Тогда $u(t)$ имеет $P+1$ непрерывную производную. Коэффициенты a_p этого ряда зависят от начального условия u_0. Ряд (4.38) справедлив в круге радиусом до ближайшей особой точки. Поэтому при $t > t_*$ он выбирает то же самое решение дифференциального уравнения (4.37), что и при $t < t_*$. Именно оно соответствует начальному условию u_0. Аналогично осуществляется продолжение решения за последующие полюсы.

Заметим, что описанная процедура продолжения применима к задачам, в которых решение имеет с полюсы либо существенно особые точки. В этом случае ряд (4.38) содержит только целые степени $(t - t_*)$. Если особая точка является точкой ветвления, то для продолжения нужно выбрать одну из ветвей. При этом (4.38) будет обобщенным степенным рядом.

Трудности. Напомним типичные трудности, встречающиеся при численном решении подобных задач. Рассмотрим любую явную схему. Пусть для простоты шаг τ является постоянным. Формально решение численное существует на каждом шаге $t_n = N\tau$. Однако вблизи первого полюса значения $u(t_n)$ быстро нарастают, и происходит переполнение. При использовании переменного шага τ_n, уменьшающегося в соответствии с ростом решения, описанная картина сохраняется.

Если взять неявную схему, то описанная картина сохраняется. Кроме того, возникает еще одна трудность. Решение на новом шаге находится из нелинейного алгебраического уравнения. Для конечного временного шага τ_n это уравнение может не иметь вещественного решения. В обоих случаях расчет положения полюса затруднен и продолжение решения по той же схеме невозможно.

Проиллюстрируем сказанное примером. Рассмотрим задачу

$$du/dt = f(u), \quad f(u) = u^2, \quad u(0) = u^0. \tag{4.39}$$

Для нее нетрудно построить точное решение

$$u = u^0/(1 - tu^0). \tag{4.40}$$

Эта задача имеет единственный полюс первого порядка при $t = t_0 = 1/u^0$. Положим для определенности $u^0 = 1$. Это решение показано на рис. 4.13 жирной линией. Оно состоит из двух ветвей гиперболы при $t < t_0$ и при $t > t_0$, разделенных вертикальной асимптотой. Этот тест достаточно прост, поскольку имеет только один полюс, а не цепочку полюсов.

Типичный расчет по явной схеме Рунге-Кутты второго порядка точности (ERK2) с аргументом t приведен на рис. 4.13. Решение быстро растет при подходе к полюсу вплоть до переполнения. При этом расчетное значение аргумента t может несколько превысить значение t_0 на величину порядка шага τ, но на самом деле расчет передает только левую ветвь решения. Перескок на вторую ветвь не происходит. Аналогичную картину дает расчет по схеме четвертого

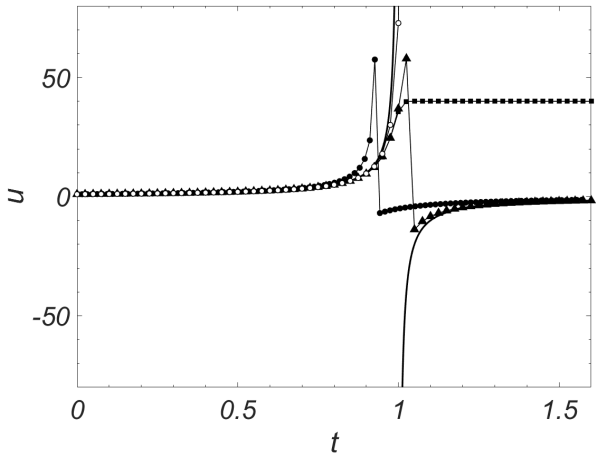

Рис. 4.13. Решения (4.39) в аргументе t по различным схемам: ∘ – ERK2, • – ROS1, ■ – CROS, ▲ – BORK2.

порядка (ERK4) и по другим явным схемам. Напомним, что расчет в аргументе l в принципе не дает перескока на вторую ветвь. Поэтому можно сказать, что явные схемы в принципе не дают автоматического продолжения решения за сингулярность.

Схемой особо высокой надежности считается явно-неявная одностадийная схема Розенброка с комплексным коэффициентом (CROS). Эта схема была подробно исследована в [125], [250]. Пример расчета по ней в аргументе t также показан на рис. 4.13. Видно, что численный расчет сначала хорошо передает левую ветвь решения. Когда расстояние до полюса становится порядка τ, численное решение формально проходит за полюс, но выходит не на вторую ветвь, а на горизонталь. Следовательно, схема CROS также не позволяет продолжать решение за сингулярность.

Рассмотрим расчет по одностадийной схеме Розенброка с коэффициентом 1 (ROS1), см. рис. 4.13. Видно, что слева от сингулярности численные решения близки к точному. Вблизи сингулярности наблюдается перескок решения с одной ветви гиперболы на другую. После этого решение идет по другой ветви гиперболы, то есть продолжается за полюс. Расчеты показали, что на различных сетках моменты перескока не одинаковы. Поэтому на первых шагах после перескока решения на всех сетках заметно отличаются от точного. При даль-

нейшем увеличении t разностное решение снова близко к точному. Таким образом, сходимость имеет место всюду, но в некоторой окрестности сингулярности она является медленной. Хотя порядок сходимости формально равен $O(\tau)$, но значения коэффициента при $t \to t_0$ неограниченно растут.

Были также проведены расчеты по следующим неявным схемам: обратная схема Эйлера точности $O(\tau)$, оптимальные обратные схемы Рунге-Кутты с двумя (BORK2) и четырьмя (BORK4) стадиями точности $O(\tau^2)$ и $O(\tau^4)$ соответственно. Оказалось, что продолжение имеет место, и результаты таковы же, как для схемы Розенброка с коэффициентом 1. При этом расчетный момент перескока с одной ветви на другую зависит от величины τ и стремится к точной асимптоте с скоростью $O(\tau)$. На рис. 4.13 приведен пример расчета по схеме BORK2.

Таким образом, прямое продолжение решения за сингулярность допускали только одностадийная схема Розенброка с коэффициентом 1 и чисто неявные обратные схемы. При этом фактическая точность составляла $O(\tau)$ независимо от теоретического порядка точности. Такая точность недостаточна для задач практики.

Заметим, что применять аргумент длина дуги при сквозном расчете через полюс не следует. Решение всегда остается на первой ветви гиперболы и в принципе не может «перепрыгнуть» на вторую, так как точная сингулярность соответствует значению аргумента $l = \infty$. Формально численное решение может оказаться правее t_0, поскольку временная компонента вычисляется с некоторой погрешностью. Максимальная расчетная длина дуги всегда конечна, поэтому численное решение никогда не достигнет сингулярности.

Замечание. В данной работе рассматривается продолжение решения за алгебраические особые точки целого порядка. Однако можно рассматривать продолжение и за полюсы дробного порядка в том случае, если по обе стороны от полюса удается однозначно определить вещественное решение. Например, такая ситуация имеет место, если полюс имеет порядок 1/3, 1/5 и т.д. Такие

задачи рассмотрены, например, в [208]. Для их численного решения с успехом применяют переход к длине дуги интегральной кривой.

4.3.2. Метод инверсной функции

Рассмотрим сначала задачу для одного ОДУ. Случай систем будет рассмотрен далее.

Введем инверсную функцию $v(t) = [u(t)]^{-1}$. Из (4.37) следует, что эта функция удовлетворяет уравнению

$$\frac{dv}{dt} = \varphi(v), \quad \varphi(v) = -v^2 f\left(\frac{1}{v}\right). \tag{4.41}$$

Это уравнение будем называть инверсным. Все полюсы функции $u(t)$ становятся нулями функции $v(t)$ и, наоборот, нули функции $u(t)$ становятся полюсами функции $v(t)$.

Выберем точку $\tilde{t} \in (0, t_*)$, такую, что на отрезке $t \in (\tilde{t}, t_*)$ точное решение $u(t)$ сохраняет знак. Для уравнения (4.41) выберем начальное условие $v(\tilde{t}) = 1/u(\tilde{t})$. Инверсное уравнение с таким начальным условием будем называть инверсной задачей. Справедливы следующие утверждения.

Теорема 12. *Пусть решение $u(t)$ задачи (4.37) имеет алгебраическую особенность в точке t_*, аналитично в проколотой окрестности этой точки и имеет $P+1$ непрерывную производную вне указанной окрестности. Особая точка t_* является полюсом первого порядка тогда и только тогда, когда $f(u)/u^2$ имеет конечный ненулевой предел при $u \to \infty$.* •

Доказательство. 1) Необходимость. Пусть u есть некоторое решение дифференциального уравнения в задаче (4.37), имеющее полюс первого порядка в точке t_*. Тогда функция $u(t)$ представима рядом Лорана (4.38). Дифференцируя его, получим ряд Лорана для сложной функции $f(u(t))$

$$f(u(t)) = \frac{du}{dt} = -\frac{a_{-1}}{(t-t_*)^2} + \sum_{p=1}^{P+1} p a_p (t-t_*)^{p-1} + o\left((t-t_*)^P\right). \tag{4.42}$$

Найдем разложение в ряд Лорана для отношения f/u^2

$$\frac{f(u(t))}{u(t)^2} = \sum_{p=0}^{P} b_p (t-t_*)^p + o\left((t-t_*)^P\right), \qquad b_0 = -\frac{1}{a_{-1}},\ b_1 = \frac{2a_0}{a_{-1}^2},\ b_2 = \frac{3}{a_{-1}^3}(a_1 a_{-1} - a_0^2), \ldots \tag{4.43}$$

Разложения (4.38), (4.42), (4.43) справедливы в некоторой малой окрестности точки t_*. Однако $t \to t_*$ как раз соответствует $u \to \infty$. Из (4.43) непосредственно видно, что в этом случае отношение $f/u^2 \to -1/a_{-1}$. Необходимость доказана.

2) Достаточность. Сделаем формальную замену неизвестной функции $u = 1/v$. Новая функция удовлетворяет инверсному уравнению (4.41), и $\varphi(v) = -v^2 f(1/v) \to C \neq 0$. Это означает, что при любом $\varepsilon > 0$ найдется такое $\delta > 0$, что при достаточно малом $|v| < \delta$ выполнено неравенство

$$C - \varepsilon < |\varphi(v)| < C + \varepsilon. \tag{4.44}$$

Выберем ε так, что $|\varepsilon| < C$. Тогда величины $C + \varepsilon$ и $C - \varepsilon$ имеют одинаковый знак, следовательно в указанной окрестности dv/dt остается знакоопределенной. Если функция $u(t)$ имеет сингулярность в точке t_*, то в этой точке $v(t)$ обращается в нуль. Из неравенств (4.44) следует, что этот нуль может быть только простым. Тогда в точке t_* функция $u(t)$ имеет полюс первого порядка. Теорема доказана. ■

Замечание 1. *Доказанная теорема близко примыкает к теории Пенлеве. В данной работе эта и последующие теоремы нужны только для обоснования предложенного метода инверсной функции.*

Замечание 2. *В доказательстве необходимости коэффициенты ряда (4.38) могут быть произвольными. Иными словами, рассматривается не одна интегральная кривая, а семейство интегральных кривых, имеющих полюс в точке t_*. Это семейство интегральных кривых можно трактовать как окрестность точного решения задачи (4.37).*

Теорема 13. *Пусть выполнены условия теоремы 12, и* $f(u)/u^2 \to C \neq 0$ *при* $u \to \infty$. *Тогда найдётся такая окрестность особой точки* t_* *и окрестность начального условия* u_0, *в которых решение* $v(t)$ *инверсной задачи существует, единственно и равномерно непрерывно зависит от начального условия* u_0. •

Доказательство. 1) Из условий теоремы следует, что в указанной окрестности $\varphi(v)$ непрерывна и ограничена, причем $\varphi(v) \to -C$ при $v \to 0$. Тогда, согласно классическим теоремам [5], решение инверсной задачи $v(t)$ существует и единственно.

2) Покажем, что оно равномерно непрерывно зависит от начального условия u_0. Расчет исходной задачи ведется до момента $\tilde{t}$, который является начальным для инверсной задачи. Согласно условиям на гладкость правой части f, решение $u(t)$ на отрезке $(0, \tilde{t})$, и в частности, его значение $u(\tilde{t})$ равномерно непрерывно зависит от u_0.

Покажем, что φ_v непрерывна. Из условий теоремы следует, что решение $u(t)$ задачи (4.37) имеет полюс первого порядка на отрезке интегрирования. Рассмотрим семейство интегральных кривых, задаваемых рядом (4.38) с произвольным коэффициентами a_p, $-1 \leq p \leq P+1$. Найдем соответствующее семейство интегральных кривых инверсного уравнения (4.41)

$$v(t) = \frac{1}{u(t)} = \sum_{p=1}^{P+1} c_p (t - t_*)^p + o\left((t - t_*)^{P+1}\right), \tag{4.45}$$

$$c_1 = \frac{1}{a_{-1}},\ c_2 = -\frac{a_0}{a_{-1}^2},\ c_3 = \frac{a_0^2 - a_1 a_{-1}}{a_{-1}^3}, \ldots \tag{4.46}$$

В силу произвольности коэффициентов a_i равенства (4.45), (4.46) задают не конкретное решение инверсной задачи, а некоторую его окрестность. Дифференцированием нетрудно найти $\varphi(v(t))$

$$\varphi(v(t)) = \frac{dv}{dt} = \sum_{p=1}^{P} p c_i (t - t_*)^{p-1} + o\left((t - t_*)^P\right) \tag{4.47}$$

Вычислим φ_v из (4.47) и (4.45) как производную параметрической функции

$$\varphi_v(v(t)) = \frac{\varphi_t(v(t))}{v_t(t)} = \left[\sum_{p=2}^{P-1} c_p p(p-1)(t-t_*)^{p-2}\right]\left[\sum_{p=2}^{P} c_p p(t-t_*)^{p-1}\right]^{-1} = \\ = \sum_{p=0}^{P-1} d_p (t-t_*)^p + o\left((t-t_*)^{P-1}\right), \tag{4.48}$$

$$d_0 = -2\frac{a_0}{a_1},\ d_1 = \frac{2}{a_{-1}^2}(a_0^2 - 3a_1a_{-1}),\ d_2 = \frac{2}{a_{-1}^3}(3a_0a_1a_{-1} - a_0^3 - 6a_2a_{-1}^2), \ldots \tag{4.49}$$

В силу оценки (4.44) $\varphi(v)$ не обращается в нуль в некоторой окрестности $v = 0$. Поэтому φ_v непрерывна, и ряд (4.48) сходится.

Таким образом, согласно классическим теоремам [5], решение инверсной задачи (4.41) равномерно непрерывно зависит от начального условия $v(\tilde{t})$. Последнее равномерно непрерывно зависит от u_0. Это завершает доказательство теоремы. ■

Следствие 1. *Построенная функция $v(t)$ является решением инверсного уравнения (4.41) как при $t < t_*$, так и при $t > t_*$. Тем самым, она однозначно определяет продолжение решения $u(t)$ за полюс по правилу $u(t) = [v(t)]^{-1}$.*

Следствие 2. *После прохождения полюса необходимо вернуться от инверсной задачи к исходной. Обозначим точку перехода через $\tilde{T}$. Примем эту точку за начальную для функции $u(t)$, зададим в ней начальное значение $u(\tilde{T}) = [v(\tilde{T})]^{-1}$ и продолжим расчет исходного уравнения при $t > \tilde{T}$. Согласно доказанным теоремам, значение $u(\tilde{T})$ зависит от u_0 равномерно непрерывно. Поэтому то же справедливо для всего решения $u(t)$ при $t > \tilde{T}$.*

Тем самым, если задача (4.37) имеет не один полюс, а последовательность полюсов, то к каждому из них применимы теоремы 12 и 13.

Следствие 3. *Из следствия 2 непосредственно вытекает, что метод инверсной функции применим к задачам с **последовательностями** полюсов первого порядка.*

Следствие 4. *Условие гладкости на функцию $f(u)$ обеспечивает применимость разностных схем порядка точности до P-го и нахождение асимптотически точного значения погрешности по Ричардсону.*

Замечание 3. *Проведенные доказательства являются локальными, то есть относятся к некоторой окрестности точки t_*. Если точка $\tilde{t}$ оказалась слишком далеко от точки t_*, то ее необходимо передвинуть в сторону t_*.*

4.3.3. Алгоритм

Процедура расчета. В численном расчете нахождение окрестностей точки t_*, указанных в теоремах 12 и 13, достаточно проблематично. Будем использовать следующий более простой алгоритм.

Выберем некоторую равномерную сетку $t_n = n\tau$, $n = 0, 1, \ldots, N$. Здесь $\tau = T/N$ – шаг сетки. Введем некоторое пороговое значение $U > 0$. Будем вести на этой сетке расчет задачи (4.37) по некоторой явной схеме до тех пор, пока для очередного сеточного значения u_n не выполнится условие $|u_n| > U$. Обозначим этот номер узла через n^*. Далее с этого момента численно решаем задачу (4.41) на той же сетке, выбирая в качестве начального условия $v_{n^*} = 1/u_{n^*}$. Во всех последующих узлах сетки восстанавливаем численное значение $u(t)$ по соотношению $u_n = 1/v_n$.

Если в ходе этого расчета v_n изменила знак, это означает переход функции $v(t)$ через нуль, что соответствует полюсу функции $u(t)$. Таким образом, мы находим решение по обе стороны полюса и можем продолжать расчет за полюс. Если при дальнейшем решении уравнения (4.41) выполнится условие $|v_n| > 1/U$, то снова возвращаемся к решению уравнения (4.37). Такой переход можно проводить неограниченное количество раз. Этот способ позволяет проводить сквозной расчет через полюс или через последовательность полюсов кратности 1. При этом не происходит никакой потери точности, поскольку уравнение (4.41) в окрестности t_* не имеет никаких особенностей.

Схемы. В любом узле сетки может произойти переход от функции $u(t)$ к инверсной функции $v(t)$. Поэтому многошаговые схемы для такого расчета непригодны. Следует использовать только одношаговые схемы. В наших расчетах хорошие результаты дали

- явные схемы Рунге-Кутты,
- явно-неявные одностадийные схемы Розенброка [49], из которых особенно надежной была вещественная схема точности $O(\tau)$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau \mathbf{w}, \quad (E - \tau \mathbf{f}_\mathbf{u})\, \mathbf{w} = \mathbf{f}(\mathbf{u}_n). \tag{4.50}$$

  Здесь E – единичная матрица. Несколько уступала ей в надежности, но имела точность $O(\tau^2)$ одностадийная схема с комплексным коэффициентом (CROS)

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau\, \mathrm{Re}\, \mathbf{w}, \quad \left(E - \frac{1+i}{2}\tau\, \mathbf{f}_\mathbf{u}\right) \mathbf{w} = \mathbf{f}(\mathbf{u}_n). \tag{4.51}$$

- диагонально-неявные схемы Рунге-Кутты BORK [60], из которых упомянем общеизвестную обратную схему Эйлера точности $O(\tau)$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau \mathbf{f}(\mathbf{u}_{n+1}). \tag{4.52}$$

  и рекурсивную схему точности $O(\tau^2)$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau \mathbf{f}\left(\mathbf{u}_{n+1} - \frac{\tau}{2}\mathbf{f}(\mathbf{u}_{n+1})\right). \tag{4.53}$$

Положения полюсов. Пусть использована численная схема точности $O(h^p)$. Найдем интервал $[t_n, t_{n+1}]$, на концах которого значения v_n и v_{n+1} имеют разные знаки. На этом интервале лежит расчетное приближение к полюсу T. В окрестности точки смены знака выберем p точек сетки t_j. Очевидно, для $p = 2$ это будут точки t_n и t_{n+1}, для $p = 4$ – точки t_{n-1}, t_n, t_{n+1}, t_{n+2}.

Рассмотрим значения v_j в выбранных узлах как аргумент, а соответствующие значения t_j – как функцию этого аргумента. Проведем ньютоновскую (или иную) интерполяцию по этим значениям v_j функции $t(v)$ и вычислим значение t_* при $v_* = 0$. Это и будет расчетное положение полюса.

Погрешность. Вычисление погрешности как традиционной нормы (C, L_2 или подобных им) от разности численного и точного решений неконструктивно на задачах с сингулярностями. Причина в том, что расчетное положение полюса заведомо отличается от точного, и поэтому погрешности вблизи полюса оказываются огромными. Для таких задач целесообразно использовать среднеквадратичный аналог метрики Хаусдорфа. Далее будем пользоваться именно этим определением.

4.3.4. Пример расчета для одного ОДУ

Автономный тест. Рассмотрим следующую задачу Коши:

$$du/dt = 1 + (u - \pi/4)^2,\ t > 0; \quad u(0) = \pi/4. \tag{4.54}$$

Приведем точное решение и его полюсы $(t_*)_m$

$$u(t) = \pi/4 + \operatorname{tg} t, \quad (t_*)_m = \pi(m - 1/2). \tag{4.55}$$

Это решение показано на рис. 4.14. Приведем также уравнение для функции $v(t)$:

$$dv/dt = -v^2 - (1 - v\pi/4)^2\,. \tag{4.56}$$

Расчеты проводились с разными константами перехода U. Оказалось, что большое значение U приводит к худшей точности. Значение $U \geq 100$ давало плохие результаты. Для иллюстративных расчетов было выбрано $U = 5$. В общем случае U есть настроечный параметр программы.

Результаты. Для численной реализации были выбраны явные схемы Рунге-Кутты второго и четвертого порядков (ERK2 и ERK4) и схема CROS (4.51). Мы пользовались их реализацией в пакете GEABORK [59]. Схемы ERK2 и CROS имеют аппроксимацию $O(\tau^2)$, а схема ERK4 – $O(\tau^4)$. Все расчеты проводились до момента $T = 10$, что требовало прохождения трех полюсов. Опишем полученные результаты.

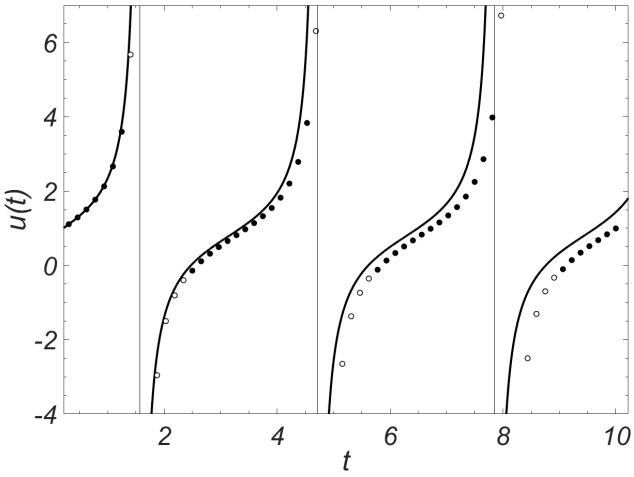

Рис. 4.14. Расчет теста (4.54) с шагом $\tau = 0.157$ по схеме ERK2. Точное решение – сплошная кривая. Точки – расчет по $u(t)$, кружки – по $v(t)$.

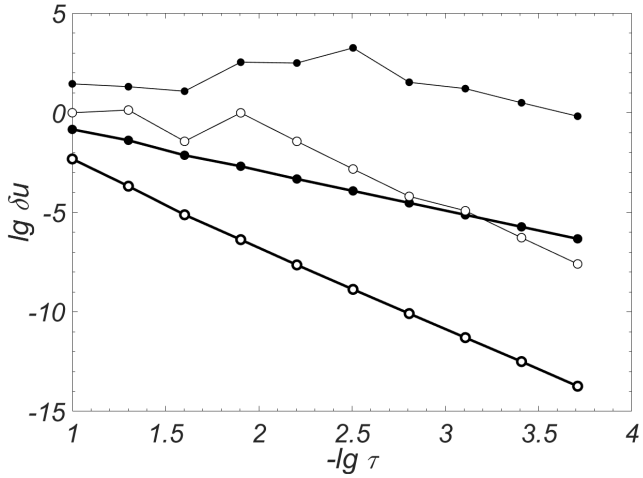

Рис. 4.15. Зависимость погрешности решения от шага, жирные линии – в метрике Хаусдорфа, тонкие – в норме L_2. Точки – схемы ERK2 и CROS, кружки – схема ERK4.

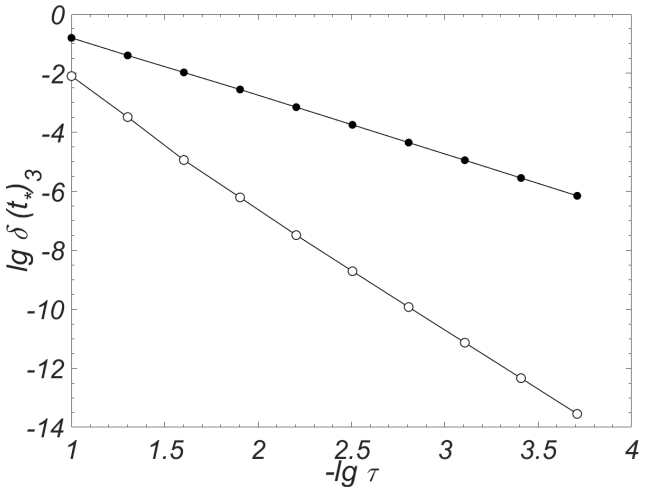

Рис. 4.16. Зависимость погрешности положения третьего полюса от шага. Точки – схемы ERK2 и CROS, кружки – схема ERK4.

На рис. 4.14 показано численное решение по схеме ERK2 с шагом $\tau = 0.157$. Шаг демонстрационного расчета выбран так, чтобы он был а) несоизмерим с расстоянием между полюсами (тогда узел сетки заведомо не попадает в полюс) и б) достаточно велик, чтобы можно было видеть отличие численного решения от точного. Видно, что расчет хорошо проходит через полюсы. Ошибка накап-

ливается с увеличением t_n, однако даже при крупном шаге можно успешно пройти много полюсов.

На рис. 4.15 для всех схем показана зависимость погрешности в метрике Хаусдорфа от величины шага. Для схем ERK2 и CROS кривые практически совпадают. График дан в двойном логарифмическом масштабе. Линии графика прямые, начиная уже со второй сетки. Следовательно, зависимость погрешности от шага является степенной. Тангенсы угла наклона этих прямых равны -2 для схем ERK2 и CROS и -4 для ERK4. Таким образом, даже на задаче с полюсами эти схемы реализуют свой теоретический порядок точности. Это свидетельствует о высокой надежности метода инверсной функции и о применимости метода Ричардсона даже на задачах с сингулярностями.

Заметим, что для такого количественного определения точности необходим использованный здесь среднеквадратичный аналог метрики Хаусдорфа. Обычные нормы L_2 или C и даже традиционная метрика Хаусдорфа (являющаяся аналогом нормы C) не дают конструктивного ответа. Для иллюстрации на рис. 4.16 показаны также погрешности в норме L_2. Сами эти погрешности намного больше погрешностей в метрике Хаусдорфа. Кривые в традиционных нормах имеют очень длинные нерегулярные участки: теоретическая сходимость начинается только при чрезмерно малом шаге.

На рис. 4.16 показана зависимость погрешности определения положения третьего полюса от шага для всех схем. График также дан в двойном логарифмическом масштабе. Картина полностью аналогична рис. 4.15: положение полюса вычисляется с точностью $O(\tau^p)$.

Обсудим некоторые следствия из этого. На грубой сетке расчетный и точный полюсы могут отстоять на несколько интервалов сетки. Если $p = 1$ (схемы первого порядка точности), то при сгущении сеток такое взаимное расположение расчетного и точного полюсов сохраняется. Но если $p \geq 2$, то после нескольких сгущений расчетный и точный полюсы попадут в один и тот же интервал сет-

ки. С этого момента на кривой погрешности в традиционной норме начинается регулярный участок.

4.3.5. Неавтономный тест

Тест (4.54) был описан автономным уравнением. Однако к такому же точному решению (4.55) приводит задача для неавтономного уравнения

$$du/dt = (u - \pi/4)\,(\operatorname{tg} t + \operatorname{ctg} t)\,. \tag{4.57}$$

В этом случае уравнение для инверсной функции имеет следующий вид:

$$dv/dt = -\left(v - v^2\pi/4\right)(\operatorname{tg} t + \operatorname{ctg} t)\,. \tag{4.58}$$

Расчет неавтономной задачи (4.57) является существенно более серьезной проблемой, чем решение автономной задачи. Поясним причину этого на данном конкретном примере. Правая часть уравнения (4.58) оказывается произведением двух сомножителей: первый зависит только от v, второй – только от t. Второй сомножитель показывает, что v должно менять знак (то есть проходить через нуль) точно при t_*. Первый же сомножитель приводит к изменению знака в точке, смещенной на один или несколько шагов τ. Поэтому вблизи t_* эти сомножители попеременно меняют знак v_n, и численное решение принимает пилообразный вид (см. рис. 4.17).

Этот эффект наблюдается на грубых сетках. Он снижает точность расчета и даже может привести к срыву расчета. Однако на достаточно подробных сетках этот эффект обычно пропадает, то есть качественный вид и точность решения оказываются хорошими. Поэтому при возникновении подобных явлений в расчетах можно порекомендовать пользователю существенно уменьшать шаг сетки. Целесообразно также увеличить разрядность вычислений.

Уравнения (4.54), (4.56), (4.58) есть уравнения Риккати. Для этого уравнения недавно был предложен [251] специализированный метод, в котором переход к новому моменту времени рассматривается как проективное преобразо-

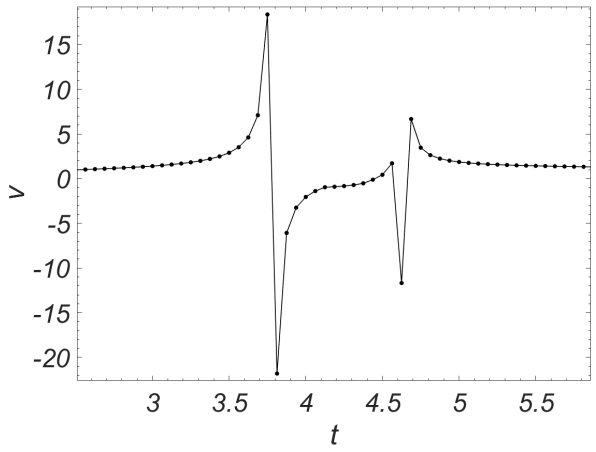

Рис. 4.17. Расчетная инверсная функция в неавтономной задаче (4.58).

вание. Этот метод успешно справляется с неавтономными уравнениями и позволяет продолжать решение за сингулярность. Описанных проблем в нем не возникает. Основным недостатком этого метода является невысокая точность $O(\tau)$.

4.3.6. Решение с ограниченной гладкостью

Выше было отмечено, что для работоспособности метода инверсной функции достаточно аналитичности решения в некоторой проколотой окрестности полюса. Вне этой окрестности решение может иметь ограниченную гладкость (но достаточную для применения той или иной разностной схемы). Этот случай не вызывает ни теоретических, ни практических затруднений. Чтобы проиллюстрировать это, рассмотрим следующую задачу:

$$\frac{du}{dt} = \begin{cases} (1-t_0)^{-2}, & 0 \leqslant t \leqslant t_0 < 1, \\ u^2, & t_0 < t < 1. \end{cases}, \quad u(0) = \frac{1-2t_0}{(1-t_0)^2}. \tag{4.59}$$

Здесь $t_0 \in [0,1]$ – некоторое фиксированное число. Нетрудно убедиться, что точное решение этой задачи имеет вид

$$u(t) = \begin{cases} (1-t_0)^{-2}(t-t_0) + (1-t_0)^{-1}, & 0 \leqslant t \leqslant t_0 < 1, \\ (1-t)^{-1}, & t_0 < t < 1. \end{cases}. \tag{4.60}$$

Решение (4.60) представляет собой линейную функцию, «склеенную» с гиперболой, и имеет вертикальную асимптоту в точке $t_* = 1$. Это решение показано на рис. 4.18. Полное время расчета T выбирается так, чтобы $T > t_*$. Очевидно, решение аналитично в круге $0 < |t - t_*| < t_* - t_0$ на комплексной плоскости. При $t = t_0$ аналитичность нарушается: в этой точке $u(t)$ непрерывно вместе с первой производной, но уже вторая производная претерпевает разрыв.

Положим $t_0 = 0.5$, $T = \pi$. Расчет этой задачи проводился по методу инверсной функции с использованием схем ERK1, ERK2 и ERK4. При указанной гладкости решения теоретическую сходимость может обеспечить только первая из них. Остальные две использовались лишь с иллюстративной целью. На рис. 4.19 показаны погрешности в зависимости от шага сетки. Масштаб графика двойной логарифмический. Погрешности решения вычислялись в среднеквадратичном аналоге метрики Хаусдорфа. Видно, что для схемы ERK1 кривая погрешности стремится к прямой с наклоном -1, что соответствует теоретическому порядку точности схемы. Аналогичное поведение имеет кривая погрешности расчета положения полюса.

Кривая погрешности решения для схемы ERK2 имеет вид «дрожащей» прямой, средний наклон которой близок к -2. Это не вполне соответствует теоретическому характеру убывания погрешности, однако даже за рамками формальной применимости эта схема обеспечивает хорошую количественную точность.

Кривая погрешности для схемы ERK4 резко отличается от прямой с наклоном -4. Это неудивительно, поскольку для применимости данной схемы решение должно иметь 5 непрерывных производных. Тем не менее, фактическая точность оказывается очень высокой (вплоть до ошибок компьютерного округления $\sim 10^{-12}$).

Таким образом, проведенные расчеты показывают, что высокие требования к гладкости решения (а именно, аналитичность) целесообразны лишь в непосредственной близости к полюсу. Вдали от полюса требования к гладкости мо-

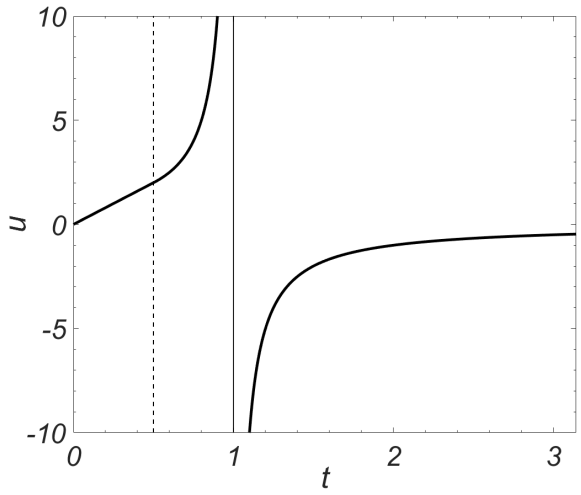

Рис. 4.18. Решение (4.60) задачи (4.59). Жирная линия – u, тонкая линия – вертикальная асимптота $t = t_*$. Пунктиром отмечена граница области аналитичности решения $t = t_0$.

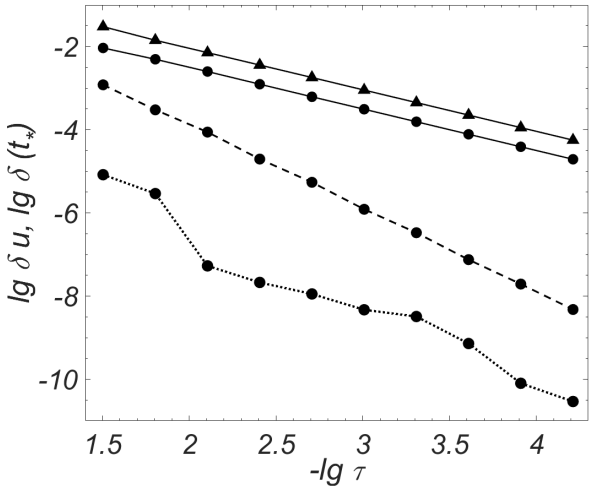

Рис. 4.19. Зависимость погрешности от шага сетки в тесте (4.59). Сплошная линия – ERK1, штриховая – ERK2, пунктир – ERK4. Круги – погрешности решения, треугольники – погрешность расчета положения полюса.

гут быть существенно ниже, и достаточно наличия нескольких непрерывных производных.

4.3.7. Системы ОДУ

Рассмотрим автономную задачу Коши для системы ОДУ с полюсами первого порядка

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}),\ t > 0,\ \mathbf{u}(0) = \mathbf{u}_0. \tag{4.61}$$

Здесь $\mathbf{u} = \{u^1, \ldots, u^J\}$, $\mathbf{f} = \{f^1, \ldots, f^J\}$. Неавтономный случай как для одного уравнения, так и для систем сводится к автономной задаче с помощью процедуры автономизации. Эта процедура будет описана далее.

Будем считать, что полюсы различных компонент не совпадают. Совпадение полюсов является вырожденным случаем, он будет рассмотрен отдельно. С формальной точки зрения, расстояние между соседними несовпадающими по-

люсами может быть любым. Однако для надежности численного расчета нужно, чтобы это расстояние было существенно больше шага сетки. Будем считать это условие выполненным (в противном случае шаг нужно уменьшить). Таким образом, полюсы различных компонент можно рассматривать последовательно один за другим.

Инверсный вектор. Пусть первым возникает полюс в компоненте u^k, соответствующий момент времени обозначим через t_*^k. В некоторой окрестности этой точки справедливы следующие разложения:

$$\begin{aligned} u^k &= \frac{a_{-1}^k}{t-t_*^k} + \sum_{p=0}^{P+1} a_p^k (t-t_*^k)^p + o\left((t-t_*^k)^{P+1}\right), \\ u^j &= \sum_{p=0}^{P+1} a_p^j (t-t_*^k)^p + o\left((t-t_*^k)^{P+1}\right), \quad 1 \le j \le J,\ j \ne k. \end{aligned} \tag{4.62}$$

Компонента, имеющая сингулярность, разлагается в ряд Лорана, а все прочие компоненты разлагаются в ряды Тейлора. Верхние пределы сумм выбраны из предположения, что правые части являются P раз непрерывно дифференцируемыми, соответственно, компоненты решения имеют $P+1$ непрерывную производную. Левую границу рассматриваемой окрестности по-прежнему будем обозначать через $\tilde{t}$.

Введем k-ю инверсную функцию $v^k(t) = [u^k(t)]^{-1}$ (ее можно рассматривать как компоненту инверсного вектора). Для нее в окрестности точки t_*^k справедливо разложение, аналогичное (4.45). Инверсная система имеет следующий вид:

$$\begin{aligned} \frac{dv^k}{dt} &= -(v^k)^2 f^k(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^J) \equiv \\ &\equiv \varphi^k(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^J), \\ \frac{du^j}{dt} &= f^j(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^J), \quad 1 \le j \le J,\ j \ne k. \end{aligned} \tag{4.63}$$

Дифференцированием рядов для $u^j(t)$, $j \ne k$ и $v^k(t)$ нетрудно получить разложения правых частей (4.63), аналогичные (4.47).

Рассмотрим инверсную систему в указанной окрестности точки t_*^k, принимая точку $\tilde{t}$ за начальную. Для компоненты v^k начальное условие имеет вид $v^k(\tilde{t}) =$

$[u^k(\tilde{t})]^{-1}$, для прочих компонент оно выписывается тривиально. Систему (4.63) вместе с указанным начальным условием будем называть инверсной задачей. Имеют место следующие теоремы.

Теорема 14. *Пусть компонента $u^k(t)$ обращается в бесконечность в точке t^k_*, аналитична в проколотой окрестности этой точки и имеет $P+1$ непрерывную производную вне указанной окрестности. Особая точка t^k_* является полюсом первого порядка тогда и только тогда, когда*

$$f^k(u^1(t), u^2(t), \dots u^J(t))/(u^k)^2$$

имеет конечный ненулевой предел при $t \to t^k_*$. •

Здесь подразумевается, что аргументами f^k являются ряды Пюизе для компонент $u^1, u^2, \dots, u^J$.

Доказательство почти дословно повторяет таковое для теоремы 1. ■

Теорема 15. *Пусть выполнены условия теоремы 3, и $f^k(u^1(t), \dots, u^J(t))/(u^k)^2$ имеет конечный ненулевой предел при $t \to t^k_*$. Тогда найдётся такая окрестность особой точки t^k_* и окрестность начального условия $\mathbf{u}_0$, в которых решение инверсной задачи существует, единственно и равномерно непрерывно зависит от начального условия $\mathbf{u}_0$.* •

Доказательство во многом аналогично доказательству теоремы 2.

1) По условию, правые части инверсной системы (4.63) непрерывны и ограничены. В силу классических теорем разрешимости задач Коши для систем ОДУ, решение инверсной системы существует и единственно.

2) Покажем, что зависимость решения инверсной задачи от $\mathbf{u}_0$ является равномерно непрерывной. Легко видеть, что начальное условие в точке $\tilde{t}$ зависит от $\mathbf{u}_0$ равномерно непрерывно. Поэтому достаточно обосновать равномерно непрерывную зависимость решения инверсной задачи от условия в точке $\tilde{t}$. Для этого достаточно показать, что компоненты якобиана системы (4.63) непрерывны.

В самом деле, в якобиан системы (4.63) входят производные вида $\partial\varphi^k/\partial v^k$, $\partial\varphi^k/\partial u^j$, $\partial f^i/\partial v^k$, $\partial f^i/\partial u^j$, здесь всюду $i, j \neq k$. Производные вида $\partial\varphi^k/\partial u^j$ и $\partial f^i/\partial u^j$ непрерывны в силу условий гладкости на правые части $\mathbf{f}$ (напомним, что φ^k зависит от u^j точно так же, как f^k).

Выражения для $\partial\varphi^k/\partial v^k$ полностью аналогичны (4.48), (4.49). Тем же способом можно найти $\partial f^i/\partial v^k$. Приведем для них несколько первых коэффициентов ряда (4.48)

$$\begin{gathered} d_0 = 2a_2^i a_{-1}^k,\ d_1 = 4a_2^i a_0^k + 6a_3^i a_{-1}^k, \\ d_2 = \frac{2}{a_{-1}^k}\left[a_2^i (a_0^k)^2 + 3a_2^i a_1^k a_{-1}^k + 6a_3^i a_0^k a_{-1}^k + 6a_4^i (a_{-1}^k)^2\right], \ldots \end{gathered} \tag{4.64}$$

Легко видеть, что производные $\partial\varphi^k/\partial v^k$ и $\partial f^i/\partial v^k$ непрерывны, поскольку φ^k не обращается в нуль в окрестности особой точки.

Таким образом, компоненты якобиана непрерывны; решение инверсной задачи зависит от $\mathbf{u}_0$ равномерно непрерывно. Теорема доказана. ■

Следствие 5. *Инверсная компонента $v^k(t)$ однозначно определяет продолжение за полюс для компоненты $u^k(t)$ и, тем самым, для всей системы (4.61).*

Следствие 6. *После прохождения полюса следует вернуться от инверсной системы к исходной. Это делается так же, как в случае одного уравнения. При этом зависимость решения за полюсом от начального условия $\mathbf{u}_0$ является равномерно непрерывной.*

Последующий полюс может быть как в той же компоненте u^k, так и в какой-то другой. В обоих случаях для него справедливы теоремы 3, 4.

Следствие 7. *Метод инверсной функции применим к системам со* ***множественными*** *полюсами первого порядка.*

Следствие 8. *Условие на гладкость $\mathbf{f}(\mathbf{u})$ позволяет применять разностные схемы порядка точности до P-го и вычислять асимптотически точное значение погрешности по Ричардсону.*

Совпадение полюсов. Рассмотрим случай, когда полюсы нескольких компонент совпадают. Пусть, для простоты, таких компонент две. Если совпадают полюсы большего числа компонент, то рассмотрение проводится полностью аналогично.

Пусть компоненты u^k и u^l имеют полюсы первого порядка, находящиеся в одной и той же точке $t_*^k = t_*^l \equiv t_*$. Пусть прочие компоненты не имеют особенностей в некоторой окрестности этой точки. Тогда в этой окрестности точки t_* уже две компоненты разлагаются в ряд Лорана:

$$\begin{aligned} u^j &= \frac{a_{-1}^j}{t-t_*} + \sum_{p=0}^{P+1} a_p^j (t-t_*)^p + o\left((t-t_*)^{P+1}\right), \quad j = k, l \\ u^j &= \sum_{p=0}^{P+1} a_p^j (t-t_*)^p + o\left((t-t_*)^{P+1}\right), \quad j \neq k, l. \end{aligned} \tag{4.65}$$

В этом случае вводится не одна инверсная компонента, а две: $v^j(t) = [u^j(t)]^{-1}$, $j = k, l$. Инверсная система имеет следующий вид:

$$\begin{aligned} \frac{dv^j}{dt} &= -(v^j)^2 f^j(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^{l-1}, [v^l]^{-1}, u^{l+1}, \ldots, u^J) \equiv \\ &\equiv \varphi^j(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^{l-1}, [v^l]^{-1}, u^{l+1}, \ldots, u^J), \ j = k, l, \\ \frac{du^j}{dt} &= f^j(u^1, u^2, \ldots, u^{k-1}, [v^k]^{-1}, u^{k+1}, \ldots, u^{l-1}, [v^l]^{-1}, u^{l+1}, \ldots, u^J), \ j \neq k, l. \end{aligned} \tag{4.66}$$

Начальное условие в точке $\tilde{t}$ выбирается очевидным образом.

Нетрудно показать, что для такой инверсной задачи справедлив аналог теоремы 4, в котором необходимо оговорить наличие конечного ненулевого предела для $f^k(\mathbf{u})/(u^k)^2$ и $f^l(\mathbf{u})/(u^l)^2$. Напомним, что это условие эквивалентны наличию полюсов первого порядка в этих компонентах.

Автономизация. Неавтономная задача сводится к автономной с помощью процедуры автономизации. Напомним эту процедуру. Исходная задача имеет следующий вид:

$$d\mathbf{u}/dt = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0. \tag{4.67}$$

Введем новую компоненту решения $u^0 \equiv t$. Для нее $f^0 = 1$. Тогда неавтономная система (4.67) переходит в автономную систему

$$d\mathbf{u}/dt = \mathbf{F}(u^0, u^1, \ldots, u^J), \quad u^0(0) = 0, \quad u^j(0) = u_0^j, \quad 1 \le j \le J, \\ F^0 = 1, \quad F^j = f^j, \quad 1 \le j \le J. \tag{4.68}$$

Если $\mathbf{f} \in \mathbb{C}^P(\mathbb{R}_J)$, то такую же гладкость имеют правые части $\mathbf{F}$. Поэтому для системы (4.68) также справедливы теоремы 3 и 4.

Алгоритм. Описанный выше алгоритм расчета легко обобщается на случай систем ОДУ. Подчеркнем, что для каждой компоненты $u^k(t)$ вектор-функции $\mathbf{u}(t)$ необходимо вводить отдельную инверсную функцию $v^k(t) = [u^k(t)]^{-1}$ (то есть компоненту инверсного вектора). Параметр A^k также следует выбирать отдельно для каждой компоненты.

4.3.8. Пример для системы ОДУ

Тест. Для иллюстрации рассмотрим следующую систему ОДУ

$$du_1/dt = u_1(u_1 + u_2), \quad du_2/dt = -u_2(u_1 + u_2), \quad u_{1,2}(0) = 1/\sqrt{2}. \tag{4.69}$$

Точное решение (4.69) имеет следующий вид:

$$u_1 = \operatorname{tg}(t - \pi/4), \quad u_2 = \operatorname{ctg}(t - \pi/4). \tag{4.70}$$

В моменты $t_k^1 = \pi/4 + \pi(k - 1/2)$, компоненты u_1 и u_2 имеют простые полюсы и простые нули соответственно. В моменты $t_k^2 = \pi/4 + \pi k$, u_1 имеет простые нули, а u_2 – простые полюсы (см. рис. 4.20).

Для разных компонент переход к соответствующей инверсной функции может происходить в различные моменты времени. Поэтому необходимо рассмотреть следующие случаи. Во-первых, пусть условие перехода выполнено только для функции u_1, то есть $|u_1| > A_1$ и $|u_2| < A_2$. Тогда от (4.69) перейдем к системе

$$dv_1/dt = -1 - v_1 u_2, \quad du_2/dt = -u_2/v_1 - u_2^2. \tag{4.71}$$

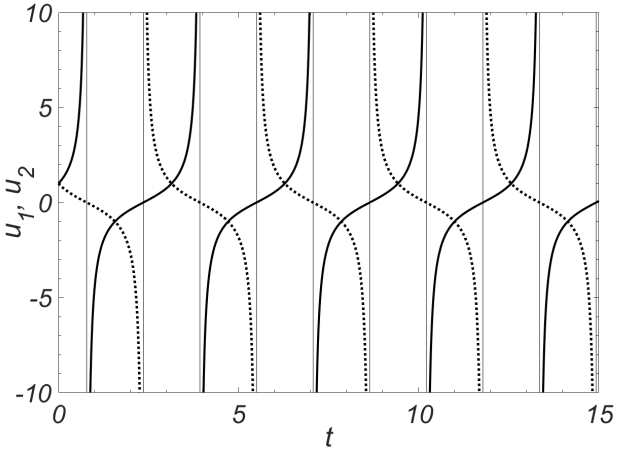

Рис. 4.20. Решение задачи (4.69). Сплошная линия – u_1, пунктир – u_2, вертикальные линии – положения полюсов.

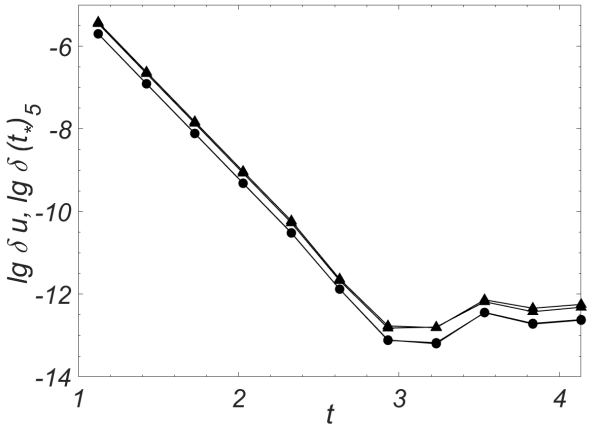

Рис. 4.21. Зависимость погрешности от шага сетки в тесте (4.69). • – решения для обеих компонент, ▲ – положение пятого полюса в обеих компонентах.

Во-вторых, если $|u_1| < A_1$ и $|u_2| > A_2$, рассмотрим систему

$$du_1/dt = u_1^2 + u_1/v_2, \quad dv_2/dt = v_2 u_1 + 1. \tag{4.72}$$

Наконец, если условия $|u_1| > A_1$, $|u_2| > A_2$ выполнены для обеих компонент, следует решать систему

$$dv_1/dt = -1 - v_1/v_2, \quad dv_2/dt = v_2/v_1 + 1. \tag{4.73}$$

Очевидно, системы (4.71) – (4.73) являются нелинейными. Они не содержат малых параметров, допускающих линеаризацию. Поэтому тест (4.69) достаточно представителен. Его важным преимуществом является то, что точное решение (4.70) выражается в элементарных функциях.

Результаты. Мы проводили расчеты по схеме ERK4. Промежуток $t \le T = 15$ включал 5 полюсов для каждой компоненты решения. На рис. 4.21 показана точность нахождения всего решения и положения пятого полюса для каждой из компонент. Обе компоненты вычислены с примерно одинаковой точностью. То же справедливо для положения всех полюсов каждой из компонент.

Видно, что даже на достаточно грубой сетке с шагом $\tau = 0.075$ получена отличная точность $\sim 3 \cdot 10^{-6}$. Точность $\sim 10^{-13}$, соответствующая ошибкам округления, достигнута при $\tau \sim 10^{-3}$. Скорость убывания ошибки при уменьшении шага сетки соответствует теоретическому четвертому порядку точности. Это подтверждает высокую надежность предложенного метода.

4.4. Множественные кратные полюсы

4.4.1. Метод обобщенной инверсной функции

Трудности. Если полюс $u(t)$ имеет кратность $q > 1$, то в точке t_* инверсная функция $v(t)$ имеет нуль кратности q. Следовательно, в этой точке обращается в нуль не только $v(t)$, но и ее производные до $(q-1)$-й включительно.

Это представляет серьезную проблему для численного решения. В самом деле, явные схемы порядка точности $p \le q$ передают p первых производных решения. В точке t_* все эти производные обращаются в нуль, поэтому при $t > t_*$ численное решение оказывается тривиальным $v \equiv 0$. Такие явные схемы позволяют дойти до полюса t_*, но не позволяют вести из него дальнейший расчет.

На практике ни один из узлов сетки не будет точно совпадать с точкой t_*. Поэтому формальное продолжение решения за полюс возможно. Но при этом значения $v(t)$ в районе полюса очень малы, и существенно возрастает роль ошибок округления. Поэтому после сквозного прохождения каждого полюса точность решения существенно ухудшается тем сильнее, чем больше кратность полюса q. Найти много полюсов высокой кратности с 64-битовыми числами не удается, и требуется существенно увеличивать разрядность вычислений, что неконструктивно. Изложим способ преодоления этой трудности.

Обобщенная инверсная функция. По скорости убывания функции $v(t)$ при приближении к полюсу можно определить кратность нуля q, который равен кратности полюса функции $u(t)$. Будем предполагать, что кратность q ближайшего полюса найдена (разумеется, разные полюсы могут иметь различную

кратность). Тогда вместо инверсной функции $v(t)$ перейдем к обобщенной инверсной функции $w(t)$. Для нечетного k она имеет следующий вид:

$$w(t) = v^{1/q}. \tag{4.74}$$

Дробная степень $1/q$ имеет q комплексных значений. При нечетном q одно из них вещественно. В формуле (4.74) подразумевается выбор вещественной ветви корня.

При четном q надо различать два случая. Если $v > 0$, имеются два вещественных значения корня, различающихся знаками. Для наших целей выбор знака безразличен. Поэтому также можно пользоваться записью (4.74). Если $v < 0$, то среди значений корней нет вещественных. Однако для наших целей можно воспользоваться следующим обобщением

$$w(t) = \mathrm{sgn}(v)|v|^{1/q}. \tag{4.75}$$

Формулой (4.75) можно пользоваться при любом знаке v. Заметим, что для нечетных q эта формула также дает правильный результат.

У обобщенной инверсной функции t_* есть простой корень. Сама функция $w(t)$ удовлетворяет уравнению

$$\frac{dw}{dt} = -\frac{1}{q} w^{1+q} f\left(\frac{1}{w^q}\right). \tag{4.76}$$

Сквозное прохождение нуля $w(t)$ является достаточно простой задачей по сравнению с задачей о сквозном прохождении кратного нуля.

Для задачи (4.76) нетрудно доказать аналоги теорем 12 и 13 для случая одного уравнения и 14 и 15 для случая системы ОДУ. При этом в теоремах 12 и 14 следует заменить f/u^2 на $f/u^{1+1/q}$. Такие теоремы также являются строгим обоснованием метода диагностики (см. п. 4.2.1).

Определение кратности полюса. Диагностика кратности полюса проводится по формулам, описанным в пункте 4.2.1, причем тогда, когда v_n становится достаточно малым. В этом случае можно принять гипотезу о том, что t_n

достаточно близко к полюсу t_*, а само поведение функции приближенно описывается формулой $v_n \approx \mathrm{const}(t_* - t_n)^q$. Одновременно с v_n продолжается расчет величины $u_n = 1/v_n$, и поведение функции u имеет вид $u_n \approx A(t_* - t_n)^{-q}$, где A – некоторая константа. По аналогии с формулой 4.14 нетрудно получить следующую формулу для q:

$$q \approx \left[1 - \frac{\ln(f_n/f_{n+1})}{\ln(v_n/v_{n+1})}\right]^{-1}. \tag{4.77}$$

Как и (4.14), соотношение (4.77) оказывается тем точнее, чем ближе расчет подошел к полюсу.

Отметим, что для применимости этих формул необходимо (но не достаточно) выполнение следующих условий:

$$v_n v_{n+1} > 0, \quad f_n f_{n+1} > 0, \quad v_n f_n < 0, \quad |v_n| > |v_{n+1}|. \tag{4.78}$$

Поэтому вычислять значения q следует только в случае выполнения этих неравенств.

В наших расчетах формулы (4.14) и (4.77) давали совпадающие результаты. Разумеется, расчетное значение q оказывалось нецелым. Но если несколько шагов подряд оно было достаточно близко к одному и тому же целому числу, то это целое число принималось за кратность полюса теста. Расчеты показали, что найденное таким образом практически всегда совпадало с реальной кратностью полюса. Поэтому данный алгоритм определения кратности был принят в наших расчетах.

4.4.2. Теоретические аспекты построения тестов

Построение представительных тестов для задач с кратными полюсами оказалось нетривиальной проблемой. Часто в литературе в качестве тестов используют задачи для трансцендент Пенлеве либо эллиптических функций. Однако точные решения этих задач не выражаются в элементарных функциях.

Хороший тест должен иметь точное решение, содержать цепочку кратных полюсов, не содержать особенности других типов. Обсудим этот вопрос подробнее.

Несингулярные особенности. Тривиальным обобщением теста (4.54) является следующая задача:

$$\frac{du}{dt} = q(u-u_0)\left[(u-u_0)^{1/q} + (u-u_0)^{-1/q}\right], \quad \text{при нечетных } q. \tag{4.79}$$

Точное решение этой задачи имеет следующий вид:

$$u(t) = u_0 + \mathrm{tg}^q t. \tag{4.80}$$

Однако у такого теста имеется специфическая трудность, которая не свойственна общим задачам с сингулярностями. В точном решении имеются точки $t = \pi m$, $m = 0, 1, \ldots$ В них производные точного решения $u^{(k)}(t) = 0$ для $k = 1, 2, \ldots, q-1$. Такие точки назовем несингулярными особенностями. Прохождение такой особенности эквивалентно расчету задачи с началом в особой точке высокого порядка. Начало расчета в особой точке является самостоятельной трудной проблемой численных методов. В данной задаче эта трудность возникает многократно, то есть при прохождении каждой несингулярной особенности. Этот вопрос выходит за рамки данной работы.

Сформулируем следующие очевидные условия.

1. Для отсутствия несингулярных особых точек необходимо, чтобы $f(u,t) \neq 0$ на интегральной кривой между любой парой соседних полюсов.
2. Для отсутствия несингулярных особых точек достаточно, чтобы $f(u,t) \neq 0$ ни в одной полосе плоскости u, t, расположенной между соседними полюсами.

Единственный полюс. Можно сконструировать тест, не содержащий несингулярной особенности. Например, это задача

$$\frac{du}{dt} = u^\nu, \quad \nu > 1, \quad u(0) = u_0. \tag{4.81}$$

Точное решение

$$u(t) = \frac{u_0}{(1 - t/t_*)^{1/(\nu-1)}}. \tag{4.82}$$

имеет полюс порядка $q = (\nu - 1)^{-1}$ в точке $t_* = u_0^{1-\nu}/(\nu - 1)$. Однако такой тест достаточно прост, так как в нем имеется только один полюс, а не цепочка полюсов.

Полюсы четного порядка. Задачи с полюсами четного порядка имеют один существенный недостаток. Они в принципе не могут быть автономными. Справедлива следующая

Теорема 16. *Если решение имеет полюс четной кратности и описывается дифференциальным уравнением первого порядка, то это уравнение не может быть автономным.* •

В самом деле, вблизи полюса четной кратности левая и правая ветви решения имеют разные знаки производных при одном и том же значении решения u. Если задача автономна, то правая часть $f(u)$ есть однозначная функция u, и она не может иметь разные знаки при одном и том же значении u. Это противоречие доказывает теорему.

Неавтономность. В разделе 4.3.5 было замечено, что для простого полюса при неавтономной записи задачи могут возникать нежелательные осцилляции численного решения вблизи полюса. Подобная трудность возникает и с кратными полюсами. Способа преодоления этой трудности найти пока не удается. Поэтому целесообразно строить тесты с автономной записью правой части.

4.4.3. Представительные тесты

Цепочка полюсов третьего порядка. Нам удалось построить тест, отвечающий сформулированным выше требованиям. В нем точное решение имеет следующий вид:

$$u = u_0 + \mathrm{tg}^3(t - t_0) + \mathrm{tg}(t - t_0). \tag{4.83}$$

Наличие члена $\mathrm{tg}^3(t-t_0)$ обеспечивает наличие полюса третьего порядка. В то же время член $\mathrm{tg}(t-t_0)$ гарантирует отсутствие несингулярным особенностей. Уравнение (4.83) является приведенным кубическим уравнением относительно $\mathrm{tg}(t-t_0)$. Из трех корней этого уравнения только один вещественный. Он имеет следующий вид:

$$\mathrm{tg}(t-t_0) = -\frac{2}{\sqrt{27}} S\,\mathrm{sh}\varphi,\ S=\mathrm{sgn}(u-u_0),\ \varphi=\frac{1}{3}\mathrm{arcsh}\left|\frac{\sqrt{27}}{2}(u-u_0)\right|. \quad (4.84)$$

Это решение можно записать и в другой форме

$$\mathrm{tg}(t-t_0) = \sqrt[3]{\frac{u}{2}+r}+\sqrt[3]{\frac{u}{2}-r},\quad r=\sqrt{\frac{u^2}{4}+\frac{1}{27}}. \quad (4.85)$$

Продифференцировав точное решение по t, получим неавтономную запись дифференциального уравнения

$$\frac{du}{dt} = \frac{1+3\mathrm{tg}^2(t-t_0)}{\cos^2(t-t_0)} = (1+3\mathrm{tg}^2(t-t_0))(1+\mathrm{tg}^2(t-t_0)). \quad (4.86)$$

Подставляя сюда $\mathrm{tg}(t-t_0)$ из (4.84) или (4.85), получим автономное уравнение вида

$$\frac{du}{dt} = \left[1+3(2/\sqrt{27}\,\mathrm{sgn}(u-u_0)\,\mathrm{sh}\varphi)^2\right]\left[1+(2/\sqrt{27}\,\mathrm{sgn}(u-u_0)\,\mathrm{sh}\varphi)^2\right] \quad (4.87)$$

или

$$\frac{du}{dt} = 3\left[(u/2+r)^{4/3}+(u/2-r)^{4/3}+1/9\right]. \quad (4.88)$$

К сожалению, построить аналогичный автономный тест для полюсов произвольного нечетного порядка q не удалось. Если взять точное решение

$$u = u_0+\mathrm{tg}(t-t_0)+\mathrm{tg}^q(t-t_0), \quad (4.89)$$

то дифференциальное уравнение в неавтономной форме выписывается без труда

$$\frac{du}{dt} = \frac{1+q\,\mathrm{tg}^{q-1}(t-t_0)}{\cos^2(t-t_0)}. \quad (4.90)$$

Однако алгебраическое уравнение относительно $\mathrm{tg}(t-t_0)$ имеет q-ю степень, а при $q\geq 5$ оно неразрешимо в радикалах.

Цепочка полюсов второго порядка. В этом случае задача может быть только неавтономной. Нам удалось построить следующий тест. Пусть точное решение имеет вид

$$u(t) = u_0 + \frac{\sin(t-t_0)}{\cos^2(t-t_0)} = \left[1 + \mathrm{tg}^2(t-t_0)\right] \sin(t-t_0). \tag{4.91}$$

оно имеет полюсы порядка $q = 2$ при $(t_*)_m = \pi/2 + \pi m$.

Дифференцируя (4.91), получаем

$$\frac{du}{dt} = \frac{1 + 2\mathrm{tg}^2(t-t_0)}{\cos(t-t_0)} = \left[1 + 2\mathrm{tg}^2(t-t_0)\right] \left[1 + \mathrm{tg}^2(t-t_0)\right] \cos(t-t_0). \tag{4.92}$$

Выразим $\mathrm{tg}(t-t_0)$ из (4.91) и подставим в (4.92)

$$\frac{du}{dt} = \left(1/2 + \sqrt{1/4 + (u-u_0)^2} + 2(u-u_0)^2\right) \cos(t-t_0). \tag{4.93}$$

Заметим, что одному и тому же точному решению могут соответствовать разные неавтономные формы записи задачи. Например, функция (4.91) является точным решением дифференциального уравнения

$$\frac{du}{dt} = \frac{1}{\cos(t-t_0)} + 2\frac{\sin^2(t-t_0)}{\cos^3(t-t_0)}. \tag{4.94}$$

Однако все попытки расчета указанного уравнения различными квадратурными формулами оказались безуспешными: счет разваливался.

Поэтому сконструированные здесь тесты (4.87), (4.88) и (4.93) представляют самостоятельную ценность. Решения имеют цепочки полюсов указанных порядков, отсутствуют особые точки других типов, и минимизировано влияние неавтономности. Эти задачи рекомендуются при тестировании других методов сквозного расчета полюсов.

4.4.4. Апробация

Цепочка полюсов третьего порядка. Из предыдущего обсуждения видно, что тест (4.87) является представительным тестом, описывающим цепочку кратных полюсов, причем полюсы являются единственными особенностями решения. В нем отсутствуют несингулярные особые точки. Поэтому апробация

метода обобщенной инверсной функции была проведена на этом тесте. В расчетах использована схема ERK4.

Были выполнены две серии расчетов теста (4.87) на последовательности равномерных сгущающихся вдвое сеток. Расчет проводился на отрезке $0 \le t \le T = 15$, содержащем пять полюсов. В первой серии во входных данных программы принудительно задавалась кратность полюса $q = 3$. В этой серии при выполнении условия $|u_n| > U$ выполнялся переход не к простой, а сразу к обобщенной инверсной функции w, соответствующей данному q. Для первой сетки задавался шаг $\tau = 0.15$. Далее сетки последовательно сгущались вдвое, пока погрешности не достигали ошибок округления. На рис. 4.22 показан график решения на самой грубой сетке. Видно, что при сквозном прохождении всех пяти полюсов третьего порядка численное решение визуально совпадает с точным. Это свидетельствует о высокой надежности метода обобщенной инверсной функции.

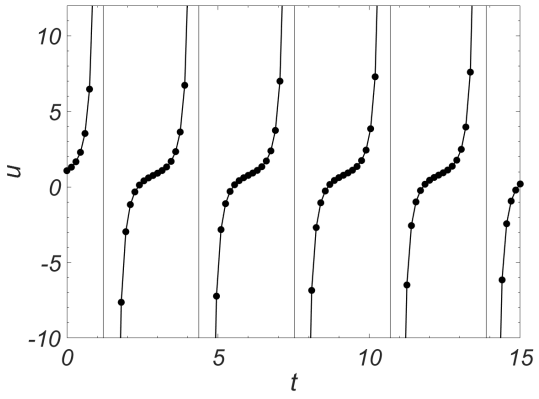

Рис. 4.22. Расчет теста (4.87) с шагом $\tau = 0.15$ по схеме ERK4. Сплошная линия – точное решение (4.83), маркеры – численное решение.

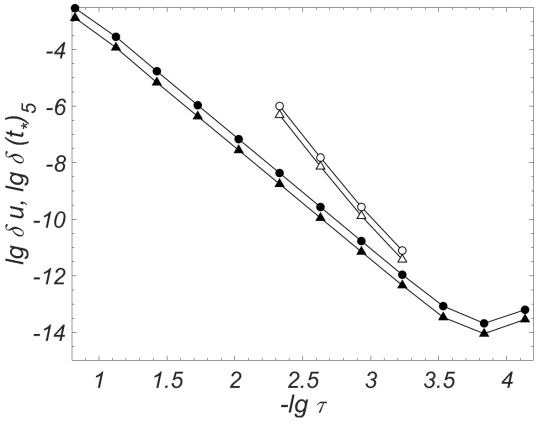

Рис. 4.23. Зависимость погрешности решения и положения пятого полюса от шага в тесте (4.87). Обозначения – см. текст.

На рис. 4.23 показана зависимость погрешностей от шага сетки в двойном логарифмическом масштабе. Каждая сетка отмечена маркером. Для удобства эти маркеры соединены между собой тонкими линиями. Линия с темными кружка-

ми показывает погрешность решения в среднеквадратичном аналоге метрики Хаусдорфа.

Видно, что начальный участок линии слегка искривлен, но быстро происходит выход на регулярный прямолинейный участок. Его наклон точно соответствует теоретическому порядку сходимости $p = 4$ схемы ERK4. Погрешность убывает до $\sim 10^{-14}$, затем перестает убывать при дальнейшем сгущении сетки. Этот предел показывает, что ошибки округления всего лишь в ~ 100 раз превышают ошибку единичного округления компьютера 10^{-16}.

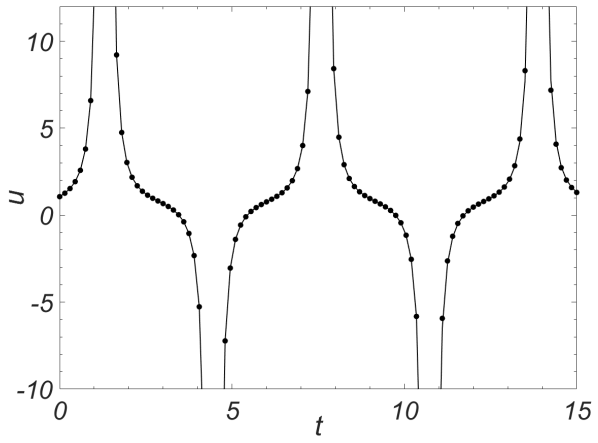

Рис. 4.24. Расчет теста (4.93) с шагом $\tau = 0.15$ по схеме ERK4. Сплошная линия – точное решение (4.83), маркеры – численное решение.

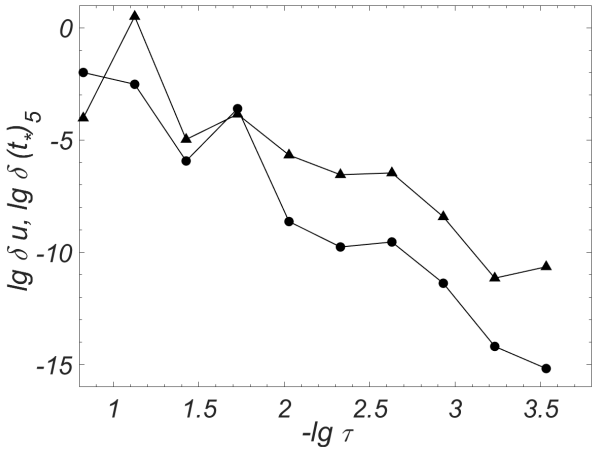

Рис. 4.25. Зависимость погрешности решения и положения пятого полюса от шага в тесте (4.93). Обозначения – см. текст.

Линия с темными треугольниками показывает погрешность определения положения самого далекого пятого полюса. Для нее также справедливо сказанное выше.

Во второй серии расчетов кратность полюса не задавалась. Программа автоматически определяла кратность по критерию (4.14). При этом сначала происходил переход от основной функции u к простой инверсной функции v. После установления кратности полюса производился переход к обобщенной инверсной функции. На практике этот критерий не срабатывал как на грубых сетках, так и

на слишком подробных. В этом случае счет разваливался. Расчеты удалось провести на сетках с числами узлов $N = 400, 800, 1600, 3200$. Погрешность решения на всем отрезке $[0, T]$ и ошибка положения пятого полюса также представлены на рис. 4.23 светлыми маркерами.

Видно, что на первой из этих четырех сеток погрешности решения и пятого полюса в ~ 100 раз превышают соответствующие значения, полученные при априорном задании правильного q. При дальнейшем сгущении сеток расхождение быстро уменьшается. Кривые стремятся к линиям, полученной при априорно заданном q. Таким образом, программа с автоматическим выбором q по точности и надежности уступает программе с априорно заданной кратностью полюса. Однако видно, что она позволяет рассчитывать цепочки полюсов, о природе которых заранее ничего неизвестно.

Сравнение двух серий расчетов показывает, что критерий определения кратности полюса существенно менее надежен, чем сам метод обобщенной инверсной функции. Разработка более надежного критерия может быть предметом дальнейшего исследования.

Цепочка полюсов второго порядка. Расчеты теста (4.93) проводились по схеме ERK4. На рис. 4.24 приведены численное решение при $\tau = 0.15$ и точное решение (обозначения соответствуют рис. 4.22). Видно, что численный расчет через 5 полюсов проходит успешно, хотя визуальное отличие к концу расчета несколько больше, чем для автономной задачи на рис. 4.22.

На рис. 4.25 показана зависимость погрешности самого решения и положения пятого полюса от шага сетки. Обозначения соответствуют рис. 4.23. Видно, расчетные точки несколько разбросаны вокруг усредненной прямой. Это и есть проявление трудностей, связанных с решением неавтономных задач. Однако средний наклон прямой соответствует теоретическому порядку точности $p = 4$, и на умеренных сетках достигается очень высокая точность, близкая к ошибкам единичного округления.

4.4.5. Пакет программ

Предлагаемый алгоритм реализован в виде прикладного пакета программ `Continuation` в среде Matlab. В нем реализован расчет скалярного ОДУ с цепочкой алгебраических особых точек целого порядка на отрезке интегрирования с автоматическим определением порядка ближайшей особенности. Он содержит подпрограмму интегрирования ОДУ по схеме Рунге-Кутты 4-го порядка точности и подпрограмму исследования ближайшей особенности и введения обобщенной инверсной функции. Ранее подобное математическое обеспечение не предлагалось. Пакет `Continuation` распространяется по свободной лицензии BSD-3-clause и доступен по ссылке
`https://github.com/ABelov91/Continuation/`.

4.5. Основные результаты главы

1. Для задач с алгебраическими особыми точками предложены новые простые и надежные способы обнаружения особенностей типа полюс, логарифмический полюс и смешанной особенности в решении систем ОДУ. Они позволяют вычислять характеристики этих особенностей с апостериорной асимптотически точной оценкой погрешности. Методика применима в аргументе длина дуги интегральной кривой, которая кардинально повышает надежность расчета. Предлагаемый подход применим и к нелинейным уравнениям в частных производных, поскольку они сводятся методом прямых к системам ОДУ огромного порядка.
2. Проведены расчеты тестовых задач с единственной сингулярностью а) для одного ОДУ с различными типами особенностей и б) задачи об S-режиме горения для квазилинейного параболического уравнения. Точное решение этих задач известно, что позволяет провести особенно тщательную апроба-

цию. Тестирование подтвердило высокую надежность предложенных методов.

3. Предложен эффективный метод решения задачи Коши для обыкновенного дифференциального уравнения с последовательностью алгебраических особых точек (как первого порядка, так и кратных). Этот метод имеет строгое обоснование. Он позволяет проводить сквозной расчет через особую точку и рассчитывать ее положение с высокой точностью. В методе используется специальный алгоритм нахождения кратности каждой особой точки. По этой кратности определяется обобщенная инверсная функция, для которой K-кратная алгебраическая особенность исходной функции является простым нулем. Расчет такого нуля не представляет трудности, поэтому предложенный метод позволяет получать высокую точность даже вблизи особых точек. После прохождения этого нуля возобновляется расчет исходной функции. Применение данного метода на последовательности алгебраических особых точек позволяет найти численное решение одновременно с апостериорной оценкой его погрешности.
4. Построены нетривиальные тестовые задачи с последовательностями алгебраических особенностей второго и третьего порядка. Эти тесты представляют самостоятельный интерес при проверке других численных методов решения задач с особыми точками. Эти тесты иллюстрируют преимущества предложенных методов.
5. Предложенные методы реализованы в виде проблемно-ориентированных комплексов программ в среде Matlab.

5. Разностные методы для одномерных уравнений Максвелла в слоистых средах

5.1. Одномерные задачи

5.1.1. Плоско-параллельная структура

1° Рассмотрим слоистую структуру, состоящую из Q изотропных плоско-параллельных пластин общей толщиной a. Ориентируем координатную ось z перпендикулярно пластинам; оси x и y ориентированы в плоскости пластин. Обозначим координаты границ слоев через $0 = \xi_0 < \xi_1 < \ldots < \xi_Q = a$. При $z < 0$ и $z > a$ расположены полубесконечные диэлектрические среды. Обозначим их диэлектрические проницаемости и магнитные восприимчивости через ε_0, μ_0 (для $z < 0$) и ε_a, μ_a (для $z > a$). Будем эти среды однородными и изотропными.

2° Обозначим через ε_q диэлектрическую проницаемость, μ_q – магнитную восприимчивость, σ_q – проводимость q-й пластины (для диэлектрических пластин $\sigma_q = 0$). Величины ε_q, μ_q, σ_q могут зависеть от частоты электромагнитной волны. Такая среда называется ***диспергирующей***.

3° Если некоторая пластина является проводящей и ее толщина пренебрежимо мала, то объемную плотность тока в ней можно представить в виде $J = j\delta(z - \xi)$, где ξ – координата пластины. Такой ток называют поверхностным. Фактически он течет по границе раздела, находящейся в точке $z = \xi$. Величина j называется поверхностной плотностью этого тока. Поверхностную проводимость обозначим через σ_q^{surf}.

Например, такая ситуация возникает, если две диэлектрические пластины разделяет тонкое металлическое напыление. Другой пример – метапленка, расположенная на подложке. Напомним, что метапленкой называют одиночный слой частиц, размер которых меньше либо сопоставим с длиной электромагнитной волны. В ряде моделей метапленку рассматривают как проводящий слой нулевой толщины [252–256].

4° В ряде актуальных задач электрическое поле $\mathbf{E}$ направлено по оси x, магнитное поле $\mathbf{H}$ – по оси y, причем амплитуды полей зависят только от координаты z. Такие задачи называют ***одномерными***.

Отметим, что в одномерных задачах объемная и поверхностная плотности зарядов равны нулю, иначе поле $\mathbf{E}$ имеет компоненту, нормальную к границе раздела. Такая задача уже не является одномерной.

Перечислим одномерные задачи, рассматриваемые в данной работе.

5.1.2. Монохроматическое излучение плоских проводников

1° Рассмотрим структуру из п. 5.1.1. Пусть часть пластин является диэлектриками, часть – проводниками или полупроводниками.

2° В проводящих пластинах текут внешние токи с объемной плотностью $\mathbf{J}_q^{\text{ext}} = \{J_q^{\text{ext}}, 0, 0\}$, направленные вдоль оси x. Будем считать, что величины J_q^{ext} зависят только от z. По границам раздела между q-й и $(q-1)$-й пластинами текут внешние поверхностные токи с поверхностными плотностями $\mathbf{j}_q^{\text{ext}} = \{j_q^{\text{ext}}, 0, 0\}$, которые также имеют только x-компоненту. Помимо внешних токов присутствуют также индуцированные токи. Для них объемная либо поверхностная плотность пропорциональна электрическому полю $J_q^{\text{ind}} = \sigma E_x$, $j_q^{\text{ind}} = \sigma^{\text{surf}} E_x$.

Поверхностные и объемные токи излучают линейно поляризованные волны в обоих направлениях оси z. При этом векторы $\mathbf{E} = \{E_x, 0, 0\}$ и $\mathbf{H} = \{0, H_y, 0\}$ направлены по осям x и y соответственно.

3° Из-за омического нагрева проводящие пластины могут становиться оптически неоднородными, то есть их показатель преломления может зависеть от координаты. Они также могут нагревать соседние диэлектрические пластины. Будем считать, что величины ε_q, μ_q, σ_q зависят от z и не зависят от x и y. Тогда и амплитуды полей зависят только от z.

4° Пусть все объемные и поверхностные токи зависят от времени как $\sim e^{-i\omega t}$, где ω – частота колебаний. Тогда поля E_x и H_y также зависят от времени

как $\sim e^{-i\omega t}$. Такую задачу называют ***монохроматической*** или ***стационарной***.

5° Описанная задача возникает в ряде актуальных приложений: излучение плоских проводников и их взаимодействие с близко расположенными слоями диэлектриков [257], проектирование печатных плат [258, 259], задачи фотовольтаики [260] и др.

6° Приведем математическую формулировку данной задачи. В каждой пластине поля подчиняются системе уравнений Максвелла. Традиционно ее рассматривают в дифференциальной форме [261–269]

$$\begin{gathered}\operatorname{rot}\mathbf{H}_q - ik_q\mathbf{D}_q - \frac{4\pi}{c}\sigma_q\mathbf{E}_q = \frac{4\pi}{c}\mathbf{J}_q^{\text{ext}},\\ \operatorname{rot}\mathbf{E}_q + ik_q\mathbf{B}_q = 0,\\ \operatorname{div}\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0,\\ \mathbf{D}_q = \varepsilon_q\mathbf{E}_q, \quad \mathbf{B}_q = \mu_q\mathbf{H}_q.\end{gathered} \tag{5.1}$$

Здесь $k_q = \omega\sqrt{\varepsilon_q\mu_q}/c$ – волновое число в q-й пластине. На границах раздела сред ставят условия сопряжения

$$\begin{gathered}\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0,\\ \mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) - \frac{4\pi}{c}\sigma_q^{\text{surf}}\mathbf{e}_z \times \mathbf{E}_q = \frac{4\pi}{c}\mathbf{j}_q^{\text{ext}},\\ \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0.\end{gathered} \tag{5.2}$$

На границах расчетной области записывают условия излучения, имитирующие уход электромагнитной волны на бесконечность

$$\frac{\partial\mathbf{E}_1}{\partial z} - ik_0\mathbf{E}_1 = 0,\ z = 0; \quad \frac{\partial\mathbf{E}_Q}{\partial z} + ik_a\mathbf{E}_Q = 0,\ z = a. \tag{5.3}$$

Здесь $k_{0,a} = \omega/c\sqrt{\varepsilon_{0,a}\mu_{0,a}}$ – волновые числа в средах, окружающих рассеиватель. Условия (5.3) являются монохроматической реализацией известных условий Мура [270].

5.1.3. Рассеяние монохроматического излучения плазмонными структурами

1° Рассмотрим структуру из п. 5.1.1. Пусть часть пластин является диэлектриками, часть – проводниками или полупроводниками.

2° Пусть на структуру с обеих сторон (т.е. из $z = -\infty$ и $z = +\infty$) нормально падают линейно поляризованные плоские монохроматические волны с частотой ω. Оси координат выберем так, чтобы векторы $\mathbf{E} = \{E_x, 0, 0\}$, $\mathbf{H} = \{0, H_y, 0\}$ были направлены по осям x, y соответственно.

Падающее излучение частично отражается от структуры и частично проходит через нее. Внутри структуры могут формироваться связанные состояния: поверхностные [271] либо таммовские [272,273] плазмон-поляритоны, моды микрополостей [274], экситоны [275] и др.

3° Внешние токи отсутствуют $\mathbf{j}^{\text{ext}} = 0$, $\mathbf{J}^{\text{ext}} = 0$. Падающее излучение индуцирует объемные токи $J_q^{\text{ind}} = \sigma_q (E_x)_q$ внутри пластин и поверхностные токи $j_q^{\text{ind}} = \sigma_q^{\text{surf}} (E_x)_q$ на границах раздела. Эти токи переизлучают линейно поляризованные волны в обоих направлениях оси z. Согласно законам отражения и преломления, частота этих волн равна частоте падающего излучения ω. Поэтому задача является монохроматической. Переизлученные волны интерферируют с падающими, отраженными и прошедшими волнами.

4° Как и в задаче 5.1.2, будем учитывать пространственную неоднородность материалов пластин, возникающую из-за нагрева токами и падающим излучением. При этом будем считать, что ε_q, μ_q, σ_q зависят только от z. В этом случае амплитуды полей также зависят только от z.

5° Описанная постановка возникает, например, в задачах плазмоники, связанных с генерацией, обнаружением и обработкой сигналов на оптических частотах вдоль границ раздела «металл-диэлектрик» [276–278].

6° Приведем математическую формулировку данной задачи. Она включает уравнения Максвелла

$$\begin{gathered}\operatorname{rot}\mathbf{H}_q - ik_q\mathbf{D}_q - \frac{4\pi}{c}\sigma_q\mathbf{E}_q = 0,\\ \operatorname{rot}\mathbf{E}_q + ik_q\mathbf{B}_q = 0,\\ \operatorname{div}\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0,\\ \mathbf{D}_q = \varepsilon_q\mathbf{E}_q, \quad \mathbf{B}_q = \mu_q\mathbf{H}_q.\end{gathered} \tag{5.4}$$

условия сопряжения

$$\begin{gathered}\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0,\\ \mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) - \frac{4\pi}{c}\sigma^{\mathrm{surf}}\mathbf{e}_z \times \mathbf{E}_q = 0,\\ \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0.\end{gathered} \tag{5.5}$$

и условия излучения на границах расчетной области

$$\frac{\partial\mathbf{E}_1}{\partial z} + ik_0\mathbf{E}_1 = 2ik_0\mathbf{E}^0,\ z = 0;\ \frac{\partial\mathbf{E}_Q}{\partial z} - ik_a\mathbf{E}_Q = 2ik_a\mathbf{E}^a e^{-ika},\ z = a. \tag{5.6}$$

Здесь $\mathbf{E}^0$, $\mathbf{E}^a$ – заданные амплитуды волн, падающих на рассеиватель из $z = -\infty$ и $z = +\infty$ соответственно.

Поясним условия (5.6). Волны, распространяющиеся в положительном и отрицательном направлении оси z, имеют вид $\mathbf{E}_+ \sim e^{i\omega t - ik_{0,a}z}$ и $\mathbf{E}_- \sim e^{i\omega t + ik_{0,a}z}$ соответственно. В общем случае по обе стороны от рассеивателя поле представимо в виде суперпозиции $\mathbf{E}_+$ и $\mathbf{E}_-$. При $z < 0$ падающей является волна $\mathbf{E}_+$, а амплитуда отраженной волны $\mathbf{E}_-$ неизвестна. Для $\mathbf{E}_-$ выполнено $\partial_z\mathbf{E}_- + ik_0\mathbf{E}_- = 0$ независимо от ее амплитуды. Для $\mathbf{E}_+$ справедливо $\partial_z\mathbf{E}_- + ik_0\mathbf{E}_- = 2ik_0\mathbf{E}^0$, это позволяет явно задать амплитуду $\mathbf{E}_+$, не затрагивая волну $\mathbf{E}_-$. При $z > a$ волны «меняются ролями»: задана падающая волна $\mathbf{E}_-$, а отраженная волна $\mathbf{E}_+$ неизвестна. Поэтому при $z = a$ граничное условие записывается аналогично таковому при $z = 0$, но со сменой знака при ik_a в левой части.

5.1.4. Рассеяние монохроматического излучения оптическими структурами

1° Пусть в задаче п. 5.1.3 все пластины являются диэлектрическими и прозрачными для падающего излучения (то есть величина $\operatorname{Im}\varepsilon \ll 1$ мала). Пусть все объемные и поверхностные токи равны нулю $J_q^{\text{ext}} = J_q^{\text{ind}} = 0$, $j_q^{\text{ext}} = j_q^{\text{ind}} = 0$.

2° Пусть на структуру из $z = -\infty$ и $z = +\infty$ нормально падают линейно поляризованные плоские монохроматические волны с частотой ω. Оси координат выберем так, чтобы векторы $\mathbf{E} = \{E_x, 0, 0\}$, $\mathbf{H} = \{0, H_y, 0\}$ были направлены по осям x, y соответственно. Падающее излучение частично отражается от структуры и частично проходит через нее.

Внутри структуры образуется стоячая волна (т.е. связанное состояние). Этот процесс является нестационарным. Во-первых, формирование такого состояния происходит не мгновенно, а в течение некоторого времени. Во-вторых, это состояние является метастабильным: с течением времени оно теряет энергию за счет излучения.

3° Пусть интенсивность падающего излучения невелика, так что нагревом пластин можно пренебречь. Тогда они являются пространственно однородными, то есть в пределах каждой пластины материальные параметры ε_q, μ_q не зависят от координат.

4° Описанную задачу в литературе традиционно называют оптической. Она встречается во многих приложениях: моделирование поляризаторов, светоделителей, ответвителей, волновых пластинок, фильтров, микролинз, микропризм, просветляющих, отражающих, рассеивающих покрытий и т.д. [268, 279–292].

5° Математическая постановка задачи имеет следующий вид:

$$\begin{gathered}\operatorname{rot}\mathbf{H}_q - ik_q\mathbf{D}_q = 0, \quad \operatorname{div}\mathbf{B}_q = 0,\\ \operatorname{rot}\mathbf{E}_q + ik_q\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0,\\ \mathbf{D}_q = \varepsilon_q\mathbf{E}_q, \quad \mathbf{B}_q = \mu_q\mathbf{H}_q.\end{gathered} \tag{5.7}$$

$$\begin{gathered}\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0,\\ \mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0.\end{gathered} \tag{5.8}$$

$$\frac{\partial\mathbf{E}_1}{\partial z} + ik_0\mathbf{E}_1 = 2ik_0\mathbf{E}^0, \ z = 0; \ \frac{\partial\mathbf{E}_Q}{\partial z} - ik_a\mathbf{E}_Q = 2ik_a\mathbf{E}^a e^{-ika}, \ z = a. \tag{5.9}$$

5.1.5. Импульсное излучение плоских проводников

1° Пусть в задаче п. 5.1.2 объемные и поверхностные токи являются финитными функциями времени, то есть представляют собой локализованные во времени импульсы. Такие импульсы содержат не одну частоту, а спектр частот. Тогда излучение таких токов также будет содержать спектр частот. Такую задачу будем называть ***немонохроматической*** или ***нестационарной***.

2° Как и в задаче п. 5.1.2, будем считать, что внешние токи $\mathbf{J}_q^{\text{ext}} = \{J_q^{\text{ext}}, 0, 0\}$, $\mathbf{j}_q^{\text{ext}} = \{j_q^{\text{ext}}, 0, 0\}$ ориентированы вдоль оси x, причем амплитуды J_q^{ext} зависят только от z. Тогда индуцированные токи $J_q^{\text{ind}} = \sigma E_x$, $j_q^{\text{ind}} = \sigma^{\text{surf}} E_x$ также направлены вдоль оси x.

Поверхностные и объемные токи излучают линейно поляризованные волны в обоих направлениях оси z. При этом векторы $\mathbf{E} = \{E_x, 0, 0\}$ и $\mathbf{H} = \{0, H_y, 0\}$ направлены по осям x и y соответственно.

3° Будем считать, что показатель преломления пластин может зависеть от координаты, но почти не зависит от времени. Иными словами, характерное время нагрева и остывания пластин существенно превосходит характерные временные масштабы задачи. По-прежнему предполагаем, что величины ε_q, μ_q, σ_q зависят только от координаты z. Тогда и амплитуды полей зависят только от z.

4° Описанная задача обобщает задачу п. 5.1.2. Она также возникает при моделировании СВЧ-техники и устройств фотовольтаики.

5° Приведем математическую постановку задачи. В (5.1) – (5.3) следует заменить множитель $(-i\omega)$ на производную ∂_t. Тогда уравнения Максвелла принимают следующий вид:

$$\begin{gathered}
\operatorname{rot}\mathbf{H}_q - \frac{1}{c}\frac{\partial \mathbf{D}_q}{\partial t} - \frac{4\pi}{c}\sigma_q \mathbf{E}_q = \frac{4\pi}{c}\mathbf{J}_q^{\text{ext}}, \\
\operatorname{rot}\mathbf{E}_q + \frac{1}{c}\frac{\partial \mathbf{B}_q}{\partial t} = 0, \\
\operatorname{div}\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0, \\
\mathbf{D}_q = \varepsilon_q \mathbf{E}_q, \quad \mathbf{B}_q = \mu_q \mathbf{H}_q.
\end{gathered} \tag{5.10}$$

В начальный момент времени поля считаются равными нулю

$$\mathbf{E}_q = 0, \quad \mathbf{H}_q = 0, \quad t = 0. \tag{5.11}$$

Условия сопряжения аналогичны (5.2), однако внешние токи могут зависеть от времени. На границах области поставим условия Мура [270]

$$\frac{\partial}{\partial t}\mathbf{E}_1 - c\frac{\partial}{\partial z}\mathbf{E}_1 = 0, \ z = 0; \quad \frac{\partial}{\partial t}\mathbf{E}_Q + c\frac{\partial}{\partial z}\mathbf{E}_Q = 0, \ z = a. \tag{5.12}$$

Условия (5.12) описывают уход импульса на бесконечность. Они основаны на том, что решения, соответствующие распространяющимся волнам, являются автомодельными $\mathbf{E}_\pm = \mathbf{E}_\pm(z \mp ct)$, т.е. зависят от комбинации переменных $z \mp ct$.

Будем считать, что среды имеют частотную дисперсию, но пространственная дисперсия пренебрежимо мала. Это означает, что величины ε_q, μ_q, σ_q, σ_q^{surf} зависят от частоты ω, но не зависят от волнового вектора.

При наличии пространственной дисперсии волновой фронт перестает быть однородным (т.е. деформируется). В этом случае задачу уже нельзя считать одномерной. Возникает деформация поляризации, которая может приводить к реализации многомодовых колебаний, волноводных режимов и др. [262]. Эти задачи выходят за рамки данной работы.

Пространственная дисперсия несущественна, если поле мало меняется на расстоянии, на котором формируется отклик среды на это поле [262]. Изменение поля происходит за счет смещения зарядов в веществе. Тем самым, мы предполагаем, что смещение зарядов за период колебаний полей мало по сравнению с длиной волны. В задачах физики плазмы это означает приближение холодной плазмы. В задачах диэлектрической фотоники и плазмоники это приближение поскольку частота колебаний поля высока. Так, типичный период колебаний в оптическом и ближнем ИК-диапазоне составляет $\sim 10^{-14}$ с. За это время свободные электроны в металле смещаются на доли ангстрема, в то время как характерная длина волны составляет сотни нанометров.

5.1.6. Рассеяние электромагнитного импульса плазмонными структурами

1° Пусть в задаче п. 5.1.3 на рассеиватель падают не монохроматические волны, а волновые пакеты

$$\begin{aligned} f^0(\zeta) &= E^0(\zeta)\exp(-i\omega^0\zeta), \quad \zeta = t - z/c, \\ f^a(\chi) &= E^a(\chi)\exp(-i\omega^0\chi), \quad \chi = t + z/c. \end{aligned} \tag{5.13}$$

с несущей частотой ω^0 и заданными огибающими $E^{0,a}(t)$. Пусть эти импульсы являются плоскими и линейно поляризованными, а огибающие – финитными. Оси координат выберем так, чтобы векторы $\mathbf{E} = \{E_x, 0, 0\}$ и $\mathbf{H} = \{0, H_y, 0\}$ были направлены по осям x и y соответственно.

Падающее излучение индуцирует объемные токи $J_q^{\text{ind}} = \sigma_q(E_x)_q$ и поверхностные токи $j_q^{\text{ind}} = \sigma_q^{\text{surf}}(E_x)_q$, которые переизлучают линейно поляризованные волны в обоих направлениях оси z. Это переизлучение содержит не одну частоту, а спектр частот.

2° Как и в задаче 5.1.5, будем считать, что показатель преломления и проводимость зависят от координаты z, но практически не меняются со временем. Тогда амплитуды полей также зависят только от z. Пусть среды имеют частот-

ную дисперсию, но эффекты пространственной дисперсии пренебрежимо малы. Условия, при которых эти предположения реализуются, указаны в п. 5.1.5.

3° Описанная задача возникает при рассмотрении сверхбыстрых процессов в плазмонных структурах; в качестве примера укажем релаксацию связанных состояний различного типа (см. [273, 293] и цитированную литературу). Эта задача является обобщением задачи п. 5.1.3.

4° Математическая постановка содержит уравнения Максвелла

$$\begin{gathered}
\operatorname{rot}\mathbf{H}_q - i\frac{1}{c}\frac{\partial\mathbf{D}_q}{\partial t} - \frac{4\pi}{c}\sigma_q\mathbf{E}_q = 0,\\
\operatorname{rot}\mathbf{E}_q + i\frac{1}{c}\frac{\partial\mathbf{B}_q}{\partial t} = 0,\\
\operatorname{div}\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0,\\
\mathbf{D}_q = \varepsilon_q * \mathbf{E}_q, \quad \mathbf{B}_q = \mu_q * \mathbf{H}_q,
\end{gathered} \tag{5.14}$$

начальные условия

$$\mathbf{E}_q = 0, \quad \mathbf{H}_q = 0, \quad t = 0. \tag{5.15}$$

Условия сопряжения совпадают с (5.5). В граничных условиях необходимо задать профиль падающего импульса. Условия Мура (5.12) не позволяют это сделать. Поэтому запишем условия излучения в виде

$$\begin{aligned}
&\frac{\partial}{\partial z}\mathbf{E}_1 - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_1 = -2\frac{\omega^0}{c}\mathbf{e}_x\frac{df^0}{d\zeta}\bigg|_{\zeta=t-z/c}, \ z = 0;\\
&\frac{\partial}{\partial z}\mathbf{E}_Q + \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_Q = -2\frac{\omega^0}{c}\mathbf{e}_x\frac{df^0}{d\chi}\bigg|_{\chi=t+z/c}, \ z = a.
\end{aligned} \tag{5.16}$$

Условия (5.16) интерпретируются аналогично (5.6).

Отметим, что в литературе в нестационарных задачах падение волны из бесконечности нередко заменяют моделью фиктивного источника (см., например, [294, 295]). Условия (5.6) и (5.16) записываются более естественно. В литературе найти их не удалось.

5.1.7. Рассеяние электромагнитного импульса оптическими структурами

1° Пусть в задаче п. 5.1.4 на рассеиватель падают плоские линейно поляризованные волновые пакеты с несущей частотой ω^0 и заданными огибающими $E^{0,a}$. Оси координат выберем так, чтобы векторы $\mathbf{E} = \{E_x, 0, 0\}$ и $\mathbf{H} = \{0, H_y, 0\}$ были направлены по осям x и y соответственно. Будем считать, что материальные параметры ε_q, μ_q постоянны в пределах каждой пластины. Имеет место частотная дисперсия, но пространственная дисперсия пренебрежимо мала.

2° Данная задача имеет большое значение для исследования динамики связанных состояний в диэлектрических структурах [296–300].

3° Математическая постановка задачи имеет следующий вид:

$$\begin{gathered}
\operatorname{rot}\mathbf{H}_q - i\frac{1}{c}\frac{\partial \mathbf{D}_q}{\partial t} - \frac{4\pi}{c}\sigma_q \mathbf{E}_q = 0, \\
\operatorname{rot}\mathbf{E}_q + i\frac{1}{c}\frac{\partial \mathbf{B}_q}{\partial t} = 0, \\
\operatorname{div}\mathbf{B}_q = 0, \quad \operatorname{div}\mathbf{D}_q = 0, \\
\mathbf{D}_q = \varepsilon_q * \mathbf{E}_q, \quad \mathbf{B}_q = \mu_q * \mathbf{H}_q,
\end{gathered} \tag{5.17}$$

$$\mathbf{E}_q = 0, \quad \mathbf{H}_q = 0, \quad t = 0, \tag{5.18}$$

$$\begin{gathered}
\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0, \\
\mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) = 0, \quad \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0.
\end{gathered} \tag{5.19}$$

$$\begin{gathered}
\frac{\partial}{\partial z}\mathbf{E}_1 - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_1 = -2\frac{\omega^0}{c}\mathbf{e}_x \frac{df^0}{d\zeta}\bigg|_{\zeta = t - z/c}, \ z = 0; \\
\frac{\partial}{\partial z}\mathbf{E}_Q + \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_Q = -2\frac{\omega^0}{c}\mathbf{e}_x \frac{df^0(\chi)}{d\chi}\bigg|_{\chi = t + z/c}, \ z = a;
\end{gathered} \tag{5.20}$$

5.1.8. Обобщенные решения

Решения задач п. 5.1.2 – 5.1.7 являются обобщенными, то есть имеют особенности. Эти особенности неподвижны, и их положения известны априори:

они соответствуют границам раздела сред. Если поверхностные токи на границах раздела равны нулю, то поля имеют слабый разрыв. Он связан со скачком фазы при отражении от каждой границы раздела. Если поверхностные токи отличны от нуля, то поле $\mathbf{H}$ имеет сильный разрыв. Его величина определяется условиями сопряжения.

5.2. Постановка задачи в интегральной форме

5.2.1. Стационарная задача

Традиционно в литературе рассматривают дифференциальную форму уравнений Максвелла. В данной работе мы будем пользоваться интегральной формой этих уравнений. Причина этого указана в п. 5.3. Напомним, что в монохроматическом случае интегральные уравнения Максвелла имеют вид

$$\int_\Gamma \mathbf{H}_q d\mathbf{l} - \frac{4\pi}{c}\int_S \sigma_q \mathbf{E}_q d\mathbf{s} + \frac{i\omega}{c}\int_S \mathbf{D}_q d\mathbf{s} = \frac{4\pi}{c}\int_S \mathbf{J}_q^{\text{ext}} d\mathbf{s}, \quad \mathbf{D}_q = \varepsilon_q \mathbf{E}_q, \tag{5.21}$$

$$\int_\Gamma \mathbf{E}_q d\mathbf{l} = \frac{i\omega}{c}\int_S \mathbf{B}_q d\mathbf{s}, \quad \mathbf{B}_q = \mu_q \mathbf{H}_q. \tag{5.22}$$

Здесь S – произвольная поверхность, ограниченная контуром Γ.

На внешних границах поставим условия излучения

$$\frac{\partial \mathbf{E}_1}{\partial z} + ik_0 \mathbf{E}_1 = 2ik_0 \mathbf{E}^0, \ z = 0; \quad \frac{\partial \mathbf{E}_Q}{\partial z} - ik_a \mathbf{E}_Q = 2ik_a \mathbf{E}^a e^{-ika}, \ z = a. \tag{5.23}$$

На границах слоев поставим условия сопряжения

$$\begin{aligned} &\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0, && \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0, \\ &\mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) - \frac{4\pi}{c}\sigma_q^{\text{surf}} \mathbf{e}_z \times \mathbf{E}_q = \frac{4\pi}{c}\mathbf{j}_q^{\text{ext}}, && \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0. \end{aligned} \tag{5.24}$$

Равенства (5.21) – (5.24) есть постановка стационарной задачи в интегральной форме.

В частности, если $E^0 = E^a = 0$; $J^{\text{ext}}, j^{\text{ext}} \neq 0$; $\sigma_q, \sigma_q^{\text{surf}} \neq 0$, то получаем задачу п. 5.1.2. При $J^{\text{ext}} = 0$, $j^{\text{ext}} = 0$; $E^0, E^a \neq 0$; $\sigma_q, \sigma_q^{\text{surf}} \neq 0$ получим задачу

п. 5.1.3. Если $J^{\text{ext}} = 0$, $j^{\text{ext}} = 0$; $E^0, E^a \neq 0$; $\sigma_q = 0$, $\sigma_q^{\text{surf}} = 0$, получаем задачу п. 5.1.4.

Таким образом, постановка (5.21) – (5.24) охватывает все одномерные монохроматические задачи, рассматриваемые в п. 5.1.

5.2.2. Нестационарная задача

Постановка задачи включает нестационарные уравнения Максвелла (5.25), (5.26), условия излучения (5.27) на границах расчетной области, условия сопряжения (5.28) на границах раздела сред, и нулевые начальные условия (5.29). Она имеет следующий вид:

$$\int_\Gamma \mathbf{H}_q d\mathbf{l} - \frac{4\pi}{c}\int_S \sigma_q \mathbf{E}_q d\mathbf{s} - \frac{1}{c}\frac{\partial}{\partial t}\int_S \mathbf{D}_q d\mathbf{s} = \frac{4\pi}{c}\int_S \mathbf{J}_q^{\text{ext}} d\mathbf{s}, \tag{5.25}$$

$$\int_\Gamma \mathbf{E}_q d\mathbf{l} = -\frac{1}{c}\frac{\partial}{\partial t}\int_S \mathbf{B}_q d\mathbf{s}, \tag{5.26}$$

$$\begin{aligned} &\frac{\partial}{\partial z}\mathbf{E}_1 - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_1 = -2\frac{\omega^0}{c}\mathbf{e}_x \left.\frac{df^0}{d\zeta}\right|_{\zeta=t-z/c}, \ z = 0; \\ &\frac{\partial}{\partial z}\mathbf{E}_Q + \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}_Q = -2\frac{\omega^0}{c}\mathbf{e}_x \left.\frac{df^a(\chi)}{d\chi}\right|_{\chi=t+z/c}, \ z = a; \end{aligned} \tag{5.27}$$

$$\begin{aligned} &\mathbf{e}_z \times (\mathbf{E}_q - \mathbf{E}_{q-1}) = 0, \qquad \mathbf{e}_z(\mathbf{D}_q - \mathbf{D}_{q-1}) = 0, \\ &\mathbf{e}_z \times (\mathbf{H}_q - \mathbf{H}_{q-1}) - \frac{4\pi}{c}\sigma_q^{\text{surf}}\mathbf{e}_z \times \mathbf{E}_q = \frac{4\pi}{c}\mathbf{j}_q^{\text{ext}}, \ \ \mathbf{e}_z(\mathbf{B}_q - \mathbf{B}_{q-1}) = 0. \end{aligned} \qquad z = \xi_q, \tag{5.28}$$

$$\mathbf{E}_q = 0, \quad \mathbf{H}_q = 0, \quad t = 0. \tag{5.29}$$

Данная постановка включает все нестационарные задачи п. 5.1. В частности, если $E^0 = E^a = 0$; $J^{\text{ext}}, j^{\text{ext}} \neq 0$; $\sigma_q, \sigma_q^{\text{surf}} \neq 0$, то (5.25) – (5.29) эквивалентна задаче п. 5.1.5. Полагая $J^{\text{ext}} = 0$, $j^{\text{ext}} = 0$; $E^0, E^a \neq 0$; $\sigma_q, \sigma_q^{\text{surf}} \neq 0$, получаем задачу п. 5.1.6. Если $J^{\text{ext}} = 0$, $j^{\text{ext}} = 0$; $E^0, E^a \neq 0$; $\sigma_q = 0$, $\sigma_q^{\text{surf}} = 0$, то получим задачу п. 5.1.7.

5.3. Бикомпактные схемы

Чтобы проводить расчеты обобщенных решений в слоистых средах, необходимо выполнение двух специфических требований: схемы должны быть бикомпактными и консервативными.

5.3.1. Консервативность

1° Как известно, уравнения математической физики выводятся из интегральных соотношений (уравнений баланса), выражающих законы сохранения для малого объема. Разностная схема должна отражать основные свойства исходной модели сплошной среды. Важнейшим свойством является выполнение разностных аналогов основных законов сохранения [116, 117, 123, 124]. Такие схемы называются ***консервативными***. Они были предложены Тихоновым и Самарским. При этом уравнение баланса для всей расчетной области должно быть алгебраическим следствием балансов в отдельных ячейках [12].

Для построения консервативных схем в [116, 117, 123] предложен интегро-интерполяционный метод. Напомним, что он основан на интегрировании исходного дифференциального уравнения по ячейке сетки, то есть на переходе от дифференциальной формы уравнения к интегральной. При этом исходное дифференциальное уравнение должно быть дивергентным. В границах раздела сред (при наличии) должны располагаться узлы сетки. Затем интегралы в полученном уравнении приближенно заменяют квадратурами.

2° В сложных задачах может быть несколько законов сохранения. Схемы, выражающие ***все*** нужные законы сохранения, называются ***полностью консервативными*** [124]. В качестве примера приведем общеизвестную систему уравнений газовой динамики. Эти уравнения выражают законы сохранения массы, импульса и энергии, которые должны выполняться одновременно. Для них построены хорошо известные полностью консервативные схемы Самарского-Попова [124].

3° Для задач электродинамики исходными законами сохранения являются закон Гаусса для электрической индукции, закон Гаусса для магнитной индукции, закон Фарадея, теорема о циркуляции магнитной индукции. Можно написать эти уравнения в дифференциальной форме, затем проинтегрировать по ячейке так же, как это делается в интегро-интерполяционном методе. Однако намного более естественно использовать сразу интегральную форму уравнений Максвелла. Это автоматически гарантирует консервативность построенной схемы. Именно поэтому мы исходим из уравнений Максвелла в интегральной форме (см. п. 5.2).

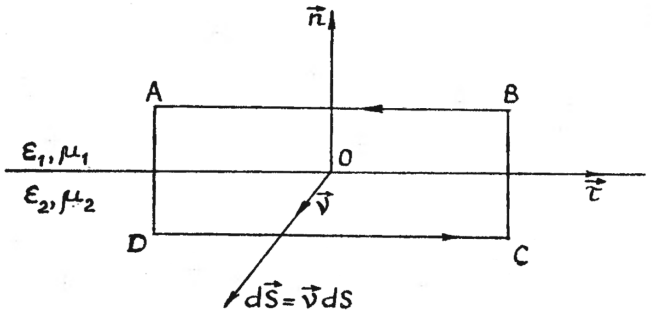

Рис. 5.1. К выводу условий сопряжения (5.24). n – нормаль к границе раздела; ν, τ – касательные векторы; контур Γ – четырехугольник $ABCD$. Иллюстрация взята из [265].

4° Как известно, условия сопряжения (5.24) являются следствием интегральных уравнений Максвелла [265, 269]. При выводе условий сопряжения выбирают такую поверхность S, что часть ее находится по одну сторону от границы раздела, часть – по другую (см. рис. 5.1).

Для этой поверхности записывают уравнения (5.21), (5.22). Далее контур стягивают в точку, лежащую на границе раздела. Это дает условия сопряжения (5.24). Таким образом, схемы в которых эти условия не учитываются, не могут быть консервативными.

Тем самым, чтобы схема для расчета электродинамики слоистых сред была полностью консервативной, она обязательно должна учитывать не только уравнения Максвелла внутри слоев, но и условия сопряжения на границах раз-

дела. В противном случае возникает сеточный дисбаланс, который влияет на решение во всей области.

Например, в схеме Йе метода FDTD это приводит к появлению нефизичных осцилляций и резкому падению точности (порядок точности ухудшается до первого) [301]. Для этой схемы доказано, что для однородной среды из разностных роторных уравнений следует разностный аналог дивергентных [295]. Однако эта схема не учитывает физически корректных условий сопряжения на границах раздела. Поэтому она не является полностью консервативной согласно указанному выше определению по Тихонову-Самарскому. То же относится и к другим конечно-разностным и конечно-элементным методам, не учитывающим условия сопряжения.

5.3.2. Бикомпактность

1° Решение задач в слоистых средах является обобщенным: на границах раздела оно испытывает излом либо разрыв. Если шаблон схемы занимает более одного шага по пространству, то граница раздела попадает внутрь шаблона, и гладкость решения внутри шаблона нарушается.

С другой стороны, традиционное исследование аппроксимации основано на разложении в ряды по степеням шага [12]. При этом предполагают определенную гладкость решения, т.е. непрерывность нескольких низших производных. Если граница раздела попадает внутрь шаблона, то это условие оказывается нарушенным. Поэтому схемы, шаблон которых занимает более одного шага сетки, не реализуют теоретический порядок точности в задачах со слоистыми средами [301]. При этом количественная погрешность нередко оказывается неприемлемо большой.

Для преодоления этих трудностей нужно, во-первых, выбирать такие сетки, в которых во всех границах раздела расположены узлы, и ни одна граница не попадает внутрь ячейки. Такие сетки называются ***специальными*** [218]. Во-

вторых, необходимо использовать двухточечные консервативные схемы. Такие схемы называются ***бикомпактными***.

Прообразом бикомпактных схем является классическая двухточечная схема Годунова [122], разработанная для газодинамической задачи о распаде разрыва. Калиткин и Корякин впервые сформулировали идею бикомпактности и построили бикомпактную схему для уравнения теплопроводности в слоистой среде в одномерном и двумерном случаях [126, 127]. Позднее Калиткин и Корякин построили бикомпактную схему для уравнения колебаний [12]. Для уравнений Максвелла бикомпактных схем ранее не существовало.

Если выбраны специальные сетки, то внутри каждого шага решение является гладким. Поэтому применимы традиционные способы исследования аппроксимации. Следовательно, бикомпактные схемы на специальных сетках обеспечивают теоретический порядок точности в любой норме даже при наличии сильных разрывов, и их погрешность быстро убывает при уменьшении шага сетки. Таким образом, бикомпактные схемы для задач в слоистых средах обладают неоспоримыми преимуществами по сравнению со схемами, в которых есть дифференцирование через границу раздела сред.

2° Общий подход к построению бикомпактных схем заключается в следующем [12]. Введем специальную сетку. Запишем закон сохранения в интегральной форме и аппроксимируем интегралы квадратурами. При этом используем двухточечный шаблон, в котором задействованы 2 соседних узла сетки. Во всех внутренних узлах сетки запишем условия сопряжения для сеточных функций, а в граничных узлах – граничные условия. Подчеркнем, что реализация условий сопряжения и граничных условий должна быть одноточечной либо двухточечной.

В силу аддитивности интегралов и соответствующих квадратур, а также благодаря явному учету условий сопряжения из выполнения законов сохранения в отдельных ячейках сетки следует выполнение тех же законов сохранения во всей расчетной области. Поэтому описанные схемы являются консерватив-

ными по Тихонову-Самарскому [123]. Именно консервативность обеспечивает сходимость к правильному обобщенному решению [117].

5.4. Стационарная бикомпактная разностная схема

5.4.1. Вывод схемы

Введем сетку $0 = z_0 < z_1 < \ldots < z_N = a$, $z_{n+1} - z_n = \Delta z_n$ так, чтобы границы слоев были узлами. Введем сеточные значения полей. В общем случае поля могут быть разрывными, поэтому в каждом внутреннем узле нужно определить левое и правое предельные значения. Для каждого шага $\Delta z_{n-1/2}$, $1 \le n \le N$ введем значения полей E_{2n-2}, H_{2n-2}, относящиеся к левой границе z_{n-1}, и значения полей E_{2n-1}, H_{2n-1}, относящиеся к правой границе z_n. Таким образом, в каждом внутреннем узле правое и левое предельные значения каждого поля могут быть неодинаковы.

2° В качестве поверхности S для (5.21) возьмем прямоугольник $z_{n-1} \le z \le z_n$, $0 \le y \le 1$ в плоскости $x = 0$. Контурный интеграл вычисляется точно

$$\int \mathbf{H}_q d\mathbf{l} = \int_0^1 (H_y)_q|_{z_{n-1}} dy - \int_0^1 (H_y)_q|_{z_n} dy = H_{2n-2} - H_{2n-1}. \tag{5.30}$$

Поверхностный интеграл в правой части (5.21) вычислим по гибридной квадратурной формуле, комбинируя правила трапеций и средних

$$\int_S \alpha \mathbf{E}_q d\mathbf{s} = \int_{z_{n-1}}^{z_n} dz \int_0^1 dy \alpha (E_x)_q \approx \frac{1}{2} \Delta z_{n-1/2} \alpha_{n-1/2} (E_{2n-1} + E_{2n-2}), \tag{5.31}$$
$$\alpha = -\frac{i\omega\varepsilon_{n-1/2}}{c} + \frac{4\pi}{c}\sigma.$$

Интеграл от плотности тока J вычислим по формуле средних. При этом значения диэлектрической проницаемости $\varepsilon_{n-1/2} \equiv \varepsilon_q\left(0.5(z_{n-1} + z_n)\right)$, магнитной восприимчивости $\mu_{n-1/2} \equiv \mu_q\left(0.5(z_{n-1} + z_n)\right)$, проводимости $\sigma_{n-1/2} \equiv \sigma_q\left(0.5(z_{n-1} + z\right.$ и объемного тока $J_{n-1/2} \equiv J_q\left(0.5(z_{n-1} + z_n)\right)$ припишем к середине ячейки. Номер пластины q ради компактности записи записи будем опускать. В результате

получим

$$H_{2n-1} - H_{2n-2} = -\alpha_{n-1/2}\Delta z_{n-1/2}(E_{2n-1} + E_{2n-2}) + \tfrac{4\pi}{c} J_{n-1/2}\Delta z_{n-1/2}, \quad 1 \le n \le N. \tag{5.32}$$

В качестве поверхности S для (5.22) выберем прямоугольник $z_{n-1} \le z \le z_n$, $0 \le x \le 1$ в плоскости $y = 0$. Аналогично предыдущему получим

$$E_{2n-1} - E_{2n-2} = \frac{i\omega}{2c}\mu_{n-1/2}\Delta z_{n-1/2}(H_{2n-1} + H_{2n-2}), \quad 1 \le n \le N. \tag{5.33}$$

3° Запишем условия сопряжения в каждом внутреннем узле

$$E_{2n} = E_{2n-1}, \quad H_{2n} - H_{2n-1} + \frac{4\pi}{c}\sigma_n^{\text{surf}} E_{2n} = -\frac{4\pi}{c} j_n, \quad 1 \le n \le N-1. \tag{5.34}$$

Здесь внешний поверхностный ток $j_n = j(z_n)$ и проводимость σ_n^{surf} относятся ко внутреннему узлу. Таким образом, каждый узел сетки может быть границей раздела сред.

4° Наконец, построим аппроксимацию условий излучения (5.23). Выразим производную $\partial_z E_q$ из дифференциального уравнения Максвелла

$$\operatorname{rot} \mathbf{E}_q = i\omega c^{-1}\mu_q \mathbf{H}_q \tag{5.35}$$

и подставим значение H_0 на левой границе и H_{2N-1} на правой. После сокращения общих множителей получим

$$\begin{gathered} \sqrt{\varepsilon_0}E_0 + \sqrt{\mu_0}H_0 = 2\sqrt{\varepsilon_0}E^0, \ z = 0 \\ \sqrt{\varepsilon_N}E_{2N-1} - \sqrt{\mu_N}H_{2N-1} = 2\sqrt{\varepsilon_N}E^a, \ z = a. \end{gathered} \tag{5.36}$$

В первом из условий (5.36) значения $\varepsilon_0 = \varepsilon(0)$ и $\mu_0 = \mu(0)$ целесообразно взять при $z = 0$. Тогда граничное условие будет передано не приближенно, а точно. Аналогично, во втором условии мы полагаем $\varepsilon_N = \varepsilon(a)$, $\mu_N = \mu(a)$.

5.4.2. Алгебраическая система

1° Равенства (5.32) – (5.36) образуют систему $4N$ алгебраических уравнений относительно $4N$ неизвестных E_{2n-2}, E_{2n-1}, H_{2n-2}, H_{2n-1}. Порядок этой

системы можно значительно понизить. Это позволяет уменьшить трудоемкость решения такой системы.

Для этого исключим из (5.34) величины $E_{2n-1} = E_{2n}$ и $H_{2n-1} = H_{2n} + 4\pi c^{-1} j_n + 4\pi c^{-1}\sigma_n^{\text{surf}} E_{2n}$ при $1 \le n \le N-1$. При этом не возникает деления на малые числа, поэтому обусловленность не ухудшается. В результате система (5.32) – (5.36) преобразуется к виду

$$\sqrt{\varepsilon_0}E_0 + \sqrt{\mu_0}H_0 = 2\sqrt{\varepsilon_0}E^0, \tag{5.37}$$

$$\begin{aligned} H_{2n} - H_{2n-2} + \alpha_{n-1/2}\Delta z_{n-1/2}(E_{2n} + E_{2n-2}) + \frac{4\pi}{c}\sigma_n^{\text{surf}} E_{2n} = \\ = \frac{4\pi}{c}(J_{n-1/2}\Delta z_{n-1/2} - j_n), \end{aligned} \tag{5.38}$$

$$\begin{aligned} E_{2n} - E_{2n-2} - \frac{i\omega}{2c}\mu_{n-1/2}\Delta z_{n-1/2}(H_{2n} + H_{2n-2}) - \frac{4\pi}{c}\sigma_n^{\text{surf}} E_{2n} = \\ = \frac{4\pi i\omega}{2c^2}\mu_{n-1/2}\Delta z_{n-1/2} j_n, \end{aligned} \tag{5.39}$$

$$H_{2N-1} - H_{2N-2} + \alpha_{N-1/2}\Delta z_{N-1/2}(E_{2N-1} + E_{2N-2}) = \frac{4\pi}{c}J_{N-1/2}\Delta z_{N-1/2}, \tag{5.40}$$

$$E_{2N-1} - E_{2N-2} - \frac{i\omega}{2c}\mu_{N-1/2}\Delta z_{N-1/2}(H_{2N-1} + H_{2N-2}) = 0, \tag{5.41}$$

$$\sqrt{\varepsilon_N}E_{2N-1} - \sqrt{\mu_N}H_{2N-1} = 2\sqrt{\varepsilon_N}E^a. \tag{5.42}$$

Напомним, что $\alpha = -i\omega\varepsilon c^{-1} + 4\pi c^{-1}\sigma$. В уравнениях (5.38), (5.39) индекс n пробегает значения от 1 до $N-1$. Таким образом, система (5.37) – (5.42) содержит $2N+2$ уравнения. Она и является разностной схемой для задачи (5.21) – (5.24).

2° Составим вектор неизвестных $\{E_0, H_0, E_2, H_2, \ldots, E_{2N-1}, H_{2N-1}\}$. Тогда система (5.37) – (5.42) имеет пятидиагональную матрицу с комплекснозначными элементами. Ее определитель содержит различные комбинации степеней шагов $\Delta z_{n-1/2}$.

Пользуясь системами символьных вычислений, мы находили определитель этой системы для небольших $N = 2, 3, \ldots$ Это позволило вычислить главный член этого определителя для произвольной неравномерной сетки и неоднородной среды

$$\text{Det} = \sqrt{\varepsilon_0\mu_N} + \sqrt{\varepsilon_N\mu_0} + ia + O(\Delta z^2). \tag{5.43}$$

Если Δz достаточно мало, то определитель отличен от нуля. Поэтому система (5.37) – (5.42) имеет и при том единственное решение.

5.4.3. Аппроксимация

Будем считать, что внутри q-й пластины функции $\varepsilon_q(z)$, $\mu_q(z)$, $J_q(z)$ являются дважды непрерывно дифференцируемыми. Тогда точное решение может иметь особенности (разрывы) только на границах раздела, а внутри пластин также имеет вторые непрерывные производные.

Напомним, что разностная схема строится на специальной сетке, узлами которой являются все границы раздела. Шаблон схемы использует только один шаг по пространству. По построению, схема консервативна. Поэтому она является бикомпактной. Для системы уравнений Максвелла такие схемы ранее не предлагались.

Таким образом, внутри шаблона решение является гладким, и для него все квадратурные формулы имеют второй порядок аппроксимации. Ошибка $O(\Delta z^2)$ вносится лишь при вычислении поверхностных интегралов типа (5.31). Контурные интегралы вычислены точно. Сеточные условия сопряжения и граничные условия также являются точными.

Введем класс функций, которые а) дважды непрерывно дифференцируемы всюду, кроме границ раздела ξ_q, и б) могут иметь разрывы в точках ξ_q, причем величины этих разрывов подчиняются условиям сопряжения (5.34). Из сказанного выше следует, что построенная схема имеет в данном классе функций аппроксимацию $O(\Delta z^2)$.

5.4.4. Устойчивость

Для простоты сначала проведем доказательство для случая непроводящей среды $\sigma = 0$. Далее обобщим результат на случай $\sigma \neq 0$.

1° Обоснуем устойчивость по граничным условиям. Пусть сначала $J_{n+1/2} = 0$, $j_n = 0$, $E^a = 0$, $E^0 \neq 0$. Рассмотрим первый шаг сетки $h_{1/2}$. Очевидно,

в амплитуда поля E в узле z_0 равна $E_0 = E^0$. Найдем, во сколько раз поле увеличивается за шаг $\Delta z_{1/2}$, предполагая, что на правой границе этого шага отражение отсутствует. В этом случае $N = 1$, и разностная схема состоит из уравнений (5.37), (5.40) – (5.42). Ее явное решение имеет вид

$$E_0 = E^0, \quad E_2 = \frac{1 - 0.5\omega\Delta z_{1/2} c^{-1}(\mathrm{n}''_{1/2} - i\mathrm{n}'_{1/2})}{1 + 0.5\omega\Delta_{1/2} c^{-1}(\mathrm{n}''_{1/2} - i\mathrm{n}'_{1/2})} E_0 \equiv A_{1/2} E_0, \qquad H_{0,2} = \sqrt{\frac{\varepsilon_{1/2}}{\mu_{1/2}}} E_{0,2}. \tag{5.44}$$

Здесь $\mathrm{n} = \mathrm{n}' + i\mathrm{n}'' = \sqrt{\varepsilon\mu}$ – показатель преломления.

Если отражение в точке z_1 есть, то амплитуда прошедшей волны заведомо не превосходит $|E_2|$. Аналогично можно оценить нарастание амплитуды поля за шаг $\Delta z_{3/2}$:

$$|E_4| \le |A_{3/2} E_2| \le |A_{3/2} A_{1/2} E^0|, \quad |H_4| = \left| \sqrt{\frac{\varepsilon_{3/2}}{\mu_{3/2}}} E_4 \right| \le \left| \sqrt{\frac{\varepsilon_{3/2}}{\mu_{3/2}}} A_{3/2} A_{1/2} E^0 \right|. \tag{5.45}$$

Продолжая эти рассуждения, найдем оценку величины поля E_{2n} во всех узлах z_n

$$|E_{2n}| \le \prod_n |A_{n+1/2}||E^0|, \quad |H_{2n}| \le \max_n \left| \sqrt{\frac{\varepsilon_{n+1/2}}{\mu_{n+1/2}}} \right| \prod_n |A_{n+1/2}||E^0|. \tag{5.46}$$

Реальные среды являются поглощающими, и для них $\mathrm{n}'' > 0$. В этом случае величина $|A_{n+1/2}|$ не превосходит единицы. Поэтому

$$|E_{2n}| \le |E^0| \tag{5.47}$$

при всех n. Отсюда следует устойчивость по граничному условию при $z = 0$. Аналогично доказывается устойчивость по граничному условию при $z = a$.

2° Докажем устойчивость по правой части. Рассмотрим объемные токи. Пусть сначала $E^0 = E^a = 0$, $j = 0$, $J_{n+1/2} \ne 0$ при некотором $n = n_0$, а все остальные значения $J_{n+1/2} = 0$. Этот ток возбуждает поля E_{2n_0}, H_{2n_0} в узле n_0 и E_{2n_0+2}, H_{2n_0+2} в узле $n_0 + 1$, которые далее распространяются в положительном и отрицательном направлении оси z. Поэтому устойчивость решения

относительно малых возмущений $J_{n_0+1/2}$ сводится к а) устойчивости значений полей E_{2n_0}, H_{2n_0}, E_{2n_0+2}, H_{2n_0+2} и б) устойчивости волн, создаваемых этими полями и распространяющихся в положительном направлении оси z.

Выразим E_{2n_0}, H_{2n_0}, E_{2n_0+2}, H_{2n_0+2} через $J_{n_0+1/2}$, считая, что за пределы данной ячейки волна уходит без отражения. Как и в предыдущем случае, разностная схема состоит из уравнений (5.37), (5.40) – (5.42). Приведем ее явное решение

$$H_{2n_0} = \frac{2\pi}{c}\frac{J_{n_0+1/2}\Delta z_{n_0+1/2}}{1-0.5i\omega\Delta z_{n_0+1/2}c^{-1}\mathrm{n}_{n_0+1/2}}, \quad H_{2n_0+2} = -H_{2n_0}, \\ E_{2n_0} = E_{2n_0+2} = H_{2n_0+2}\sqrt{\frac{\mu_{n_0+1/2}}{\varepsilon_{n_0+1/2}}}. \tag{5.48}$$

Эти значения можно рассматривать как источники волн, распространяющихся в положительном и отрицательном направлении оси z. Для этих волн справедлива оценка $|E_{2n}| \le |E_{2n_0}|$. Тем самым, для всех n имеют место неравенства

$$|H_{2n}| \le \frac{2\pi}{c}\frac{|J_{n_0+1/2}|\Delta z_{n_0+1/2}}{|1-0.5i\omega\Delta z_{n_0+1/2}c^{-1}\mathrm{n}_{n_0+1/2}|} \equiv B, \quad |E_{2n}| \le B\max_n\left|\sqrt{\frac{\mu_{n+1/2}}{\varepsilon_{n+1/2}}}\right|. \tag{5.49}$$

Если объемные токи присутствуют в нескольких ячейках, то в выражениях (5.49) нужно провести суммирование по соответствующим n_0

$$|H_{2n}| \le \frac{2\pi a}{c}\max_n|J_{n+1/2}| \equiv C, \quad |E_{2n}| \le C\max_n\left|\sqrt{\frac{\mu_{n+1/2}}{\varepsilon_{n+1/2}}}\right|. \tag{5.50}$$

Для реальных физических сред $\mathrm{n}' \ge 1$, $\mathrm{n}'' \le 0$; поэтому знаменатель (5.49) можно оценить снизу единицей. Из (5.50) следует устойчивость решения относительно малых возмущений объемных токов.

3° Полностью аналогично обосновывается устойчивость решения относительно малых возмущений поверхностных токов. Пусть сначала $E^0 = E^a = 0$, $J = 0$, $j_{n_0} \ne 0$, $j_n = 0$ при $n \ne n_0$. Этот ток возбуждает поля E_{2n_0}, H_{2n_0} в узле n_0, которые далее распространяются в положительном и отрицательном направлении оси z. Для установления связи между указанными значениями полей и током j_{n_0} рассмотрим среду, состоящую из двух ячеек $\Delta z_{n_0-1/2}$ и $\Delta z_{n_0+1/2}$.

Подчеркнем, что левое и правое предельные значения поля H в узле n_0 будут различны: $H_{2n_0-1} \neq H_{2n_0}$. При этом левое и правое предельные значения поля $E = E_{2n_0}$ одинаковы. Будем считать, что за пределы ячеек волна уходит без отражения. Разностная схема содержит граничные условия (5.37), (5.42), одну пару уравнений (5.38), (5.39) и пару уравнений (5.40), (5.41). После несложных, но громоздких вычислений получим

$$H_{2n_0} = \frac{2\pi}{c}\mu_{n_0+1/2}j_{n_0}, \quad H_{2n_0-1} = \frac{2\pi}{c}(\mu_{n_0+1/2}+2)j_{n_0}, \quad E_{2n_0} = \sqrt{\frac{\mu_{n_0+1/2}}{\varepsilon_{n_0+1/2}}}H_{2n_0}. \tag{5.51}$$

Созданные этими источниками волны распространяются в положительном и отрицательном направлении оси z. Поэтому для правого и левого предельных значений полей во всех узлах справедливы оценки

$$|H_{2n}|, |H_{2n-1}| \leq \frac{2\pi}{c}(\max_n |\mu_{n+1/2}| + 2)|j_{n_0}| \equiv D, \quad |E_{2n}| \leq \max_n \left|\sqrt{\frac{\mu_{n_0+1/2}}{\varepsilon_{n_0+1/2}}}\right| D \tag{5.52}$$

Если поверхностный ток имеется на нескольких границах раздела, то оценки (5.52) нужно просуммировать по соответствующим n_0. Заметим, что эта сумма будет включать фиксированное число слагаемых, не зависящее от числа шагов сетки. Из оценки (5.52) следует устойчивость решения относительно малых возмущений поверхностных токов j.

4° Чтобы учесть проводимость, во всех предыдущих формулах достаточно сделать замену $-i\omega\varepsilon/c \to -i\omega\varepsilon/c + 4\pi\sigma/c$. При этом доказательство устойчивости проводится полностью аналогично, но выкладки оказываются более громоздкими.

5° Таким образом, проведено полное обоснование устойчивости разностной схемы по граничным условиям и правой части. Оно справедливо для произвольных неравномерных сеток и неоднородных сред. Отсюда и из однозначной разрешимости системы уравнений (5.37) – (5.42) следует корректность предложенной разностной схемы.

5.4.5. Сходимость

1° Согласно известным теоремам теории разностных схем, из корректности разностной схемы и установленной выше аппроксимации вытекает сходимость. Напомним, что аппроксимация имеет место на классе функций, имеющих разрывы на границах раздела сред, причем величина разрыва должна подчиняться условию сопряжения. При этом необходимо использовать специальные сетки. Поэтому справедлива

Теорема 17. *Разностная схема (5.37) – (5.42) на специальной сетке сходится со вторым порядком точности.* •

2° Для предложенной схемы применим метод сгущения сеток и оценки точности по правилу Ричардсона [180, 186, 302]. В литературе предлагались разные многосеточные методы (т.н. sub-gridding techniques), см., например, [303–305]. Эти методы используют локальное сгущение для построения адаптивных сеток. Однако эти методы не дают гарантированных оценок точности. Напротив, глобальное сгущение сеток по методу Ричардсона дает асимптотически точное значение погрешности. Этот метод существенно повышает надежность расчета. Он имеет строгое обоснование [110].

3° Схема фактически построена интегро-интерполяционным методом. Поэтому для сред с непрерывно изменяющимися свойствами консервативность схемы следует из аддитивности интегралов. При наличии границ раздела консервативность обеспечивается явным учетом условий сопряжения. Таким образом, схема является полностью консервативной в смысле определения Тихонова-Самарского.

Бикомпактность и полная консервативность обеспечивают сходимость численного решения к точному даже если последнее имеет сильные или слабые разрывы.

5.5. Решение системы разностных уравнений

Разностная схема (5.37) – (5.42) есть система большого количества линейных алгебраических уравнений. В общем случае для ее решения целесообразно применять метод Гаусса. Однако есть ряд важных частных случаев, когда эта система допускает явное решение в конечном виде. Подобные случаи являются уникальными. По экономичности такие решения превосходят даже явные схемы бегущего счета (см. [12]). Насколько нам известно, единственное ранее известное решение такого класса было построено Калиткиным для уравнения Ван-Хопфа. К сожалению, оно не было опубликовано в открытой печати.

Приведенное далее решение разностной схемы (5.37) – (5.42) справедливо для однородных диэлектрических сред $\varepsilon = \text{const}$, $\mu = \text{const}$, $\sigma = 0$. Оно легко обобщается на проводящие среды, у которых $\sigma = \text{const} \neq 0$. Для этого достаточно сделать замену $-i\omega\varepsilon/c \to -i\omega\varepsilon/c + 4\pi\sigma/c$.

Сначала рассмотрим ряд простых частных задач и установим общие закономерности поведения решения, затем воспользуемся принципом суперпозиции.

5.5.1. Однородная среда

1° Рассмотрим задачу об однородной среде $\varepsilon = \text{const}$, $\mu = \text{const}$, $\varepsilon \neq \mu$. Пусть слева падает волна с амплитудой E^0, а волна справа отсутствует $E^a = 0$. Пусть сначала сетка содержит $N = 1$ шаг $\Delta z_{1/2}$. Тогда разностная схема содержит 4 уравнения

$$\begin{aligned} &\sqrt{\mu}H_0 + \sqrt{\varepsilon}E_0 = 2\sqrt{\varepsilon}E^0, && E_2 - E_0 - \beta_{1/2}\mu(H_2 + H_0) = 0, \\ &H_2 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = 0, && \sqrt{\mu}H_2 - \sqrt{\varepsilon}E_2 = 0. \end{aligned} \tag{5.53}$$

Здесь введено обозначение

$$\beta_{1/2} = 0.5i\frac{\omega}{c}\Delta z_{1/2}. \tag{5.54}$$

Из условий сопряжения следует, что в данной задаче поля E и H непрерывны в узлах. Поэтому можно использовать только значения полей E_{2n}, H_{2n} с четными

индексами, считая, что значения полей с нечетными индексами исключены из системы.

Решение системы (5.53) имеет следующий вид:

$$E_0 = E^0, \quad H_0 = E^0\sqrt{\frac{\varepsilon}{\mu}}, \quad E_2 = E^0\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}, \quad H_2 = E^0\sqrt{\frac{\varepsilon}{\mu}}\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}. \quad (5.55)$$

Здесь $\mathrm{n} = \sqrt{\varepsilon\mu}$ – показатель преломления. Легко видеть, что

$$\frac{E_2}{E_0} = \frac{H_2}{H_0} = \frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}} \equiv A_{1/2}. \quad (5.56)$$

Величину $\ln A_{1/2}$ можно трактовать как набег фазы волны при переходе от точки 0 к точке $\Delta z_{1/2}$. Нетрудно убедиться, что она совпадает с точным значением $2\beta_{1/2}\mathrm{n}\Delta z_{1/2}$ с точностью до членов $O(\Delta z^2)$.

2° Пусть теперь сетка содержит $N = 2$ ячейки $\Delta z_{1/2}$, $\Delta z_{3/2}$. Запишем соответствующую систему разностных уравнений

$$\begin{aligned} &\sqrt{\mu}H_0 + \sqrt{\varepsilon}E_0 = 2\sqrt{\varepsilon}E^0, && E_2 - E_0 - \beta_{1/2}\mu(H_2 + H_0) = 0,\\ &H_2 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = 0, && E_4 - E_2 - \beta_{3/2}\mu(H_4 + H_2) = 0,\\ &H_4 - H_2 - \beta_{3/2}\varepsilon(E_4 + E_2) = 0, && \sqrt{\mu}H_4 - \sqrt{\varepsilon}E_4 = 0. \end{aligned} \quad (5.57)$$

Величина $\beta_{3/2}$ определяется аналогично (5.54).

Решение системы (5.57) имеет следующий вид. Компоненты E_0, H_0, E_2, H_2 совпадают с (5.55), компоненты E_4, H_4 равны

$$E_4 = E^0\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}\frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, \quad H_4 = E^0\sqrt{\frac{\varepsilon}{\mu}}\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}\frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}. \quad (5.58)$$

Имеют место соотношения

$$\frac{E_4}{E_2} = \frac{H_4}{H_2} = \frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}} \equiv A_{3/2} \quad (5.59)$$

Таким образом, решение задачи (5.57) можно представить в виде

$$\begin{aligned} &E_0 = E^0, && E_2 = A_{1/2}E_0, && E_4 = A_{1/2}A_{3/2}E_0;\\ &H_0 = \sqrt{\varepsilon/\mu}E^0, && H_2 = A_{1/2}H_0, && H_4 = A_{1/2}A_{3/2}H_0. \end{aligned} \quad (5.60)$$

Формулы (5.60) имеют простую физическую интерпретацию. Волна распространяется сначала на шаг $\Delta z_{1/2}$, затем на шаг $\Delta z_{3/2}$. Материальные параметры в

ячейках $\Delta_{1/2}$ и $\Delta_{3/2}$ одинаковы, поэтому переотражений от границ ячеек не возникает. Значения полей E_2, H_2 в узле $n = 1$ можно трактовать как граничные условия для области, состоящей из шага $\Delta z_{3/2}$. Поэтому к ней можно снова применить формулы (5.55), заменив $E^0 \to E_2$. Аналогично можно поступить в случае сетки, содержащей $N = 3$ интервала и т.д.

3° В (5.60) легко просматривается закономерность, которая позволяет выписать решение для сетки с произвольным числом шагов. Это решение имеет следующий вид:

$$E_{2n} = E^0 \prod_{k=1}^{n} \frac{1 + \beta_{k-1/2}\mathrm{n}}{1 - \beta_{k-1/2}\mathrm{n}}, \quad H_{2n} = \sqrt{\frac{\varepsilon}{\mu}} E_{2n}. \tag{5.61}$$

Формулу (5.61) нетрудно доказать по индукции.

4° Полностью аналогично строится решение для случая, когда волна приходит справа $E^0 = 0$, $E^a \neq 0$. Тогда в (5.61) нужно заменить $E^0 \to E^a$ и изменить знак у H

$$E_{2n} = E^a \prod_{k=1}^{n} \frac{1 + \beta_{k-1/2}\mathrm{n}}{1 - \beta_{k-1/2}\mathrm{n}}, \quad H_{2n} = -\sqrt{\frac{\varepsilon}{\mu}} E_{2n}. \tag{5.62}$$

По принципу суперпозиции, если оба граничных условия являются ненулевыми, то нужно взять сумму (5.61), (5.62).

5.5.2. Объемные токи

1° Рассмотрим задачу об однородной среде, в которой течет объемный ток. Пусть сетка содержит $N = 1$ ячейку, в которой течет ток с плотностью $J_{1/2}$. Пусть волны, падающие из бесконечности, отсутствуют $E^0 = E^a = 0$. Разностная схема имеет вид

$$\begin{aligned} &\sqrt{\varepsilon} E_0 + \sqrt{\mu} H_0 = 0, && E_2 - E_0 - \beta_{1/2}\mu(H_2 + H_0) = 0, \\ &H_2 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = \frac{4\pi}{c} J_{1/2} \Delta z_{1/2}, && \sqrt{\varepsilon} E_2 - \sqrt{\mu} H_2 = 0. \end{aligned} \tag{5.63}$$

Решение этой разностной задачи имеет вид

$$E_0 = E_2 = \frac{2\pi}{c} \sqrt{\frac{\mu}{\varepsilon}} \frac{J_{1/2} \Delta z_{1/2}}{1 - \beta_{1/2}\mathrm{n}} \equiv \mathcal{E}^J_{1/2}, \quad H_0 = -H_2 = -\frac{2\pi}{c} \frac{J_{1/2} \Delta z_{1/2}}{1 - \beta_{1/2}\mathrm{n}}. \tag{5.64}$$

Имеют место закономерности: $H_2 = -\sqrt{\varepsilon/\mu}E_2$, $H_0 = \sqrt{\varepsilon/\mu}E_0$. Величина $\mathcal{E}^J_{1/2}$ есть функция точечного источника, соответствующего объемным токам.

2° Пусть теперь сетка содержит $N = 3$ ячейки, причем плотность тока $J_{3/2} \neq 0$, относящаяся к среднему шагу, отлична от нуля, а токи в крайних шагах $J_{1/2} = J_{5/2} = 0$ равны нулю. Запишем систему разностных уравнений

$$\begin{aligned}
&\sqrt{\varepsilon}E_0 + \sqrt{\mu}H_0 = 0, && E_2 - E_0 - \beta_{1/2}\mu(H_2 + H_0) = 0,\\
&H_2 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = 0, && E_4 - E_2 - \beta_{3/2}\mu(H_4 + H_2) = 0,\\
&H_4 - H_2 - \beta_{3/2}\varepsilon(E_4 + E_2) = \frac{4\pi}{c}J_{3/2}\Delta z_{3/2}, && E_6 - E_4 - \beta_{5/2}\mu(H_6 + H_4) = 0,\\
&H_6 - H_4 - \beta_{5/2}\varepsilon(E_6 + E_4) = 0, && \sqrt{\varepsilon}E_6 - \sqrt{\mu}H_6 = 0.
\end{aligned} \tag{5.65}$$

Решение этой разностной задачи имеет вид

$$\begin{gathered}
E_0 = \frac{2\pi}{c}\sqrt{\frac{\mu}{\varepsilon}}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}}\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}, \quad E_2 = E_4 = \frac{2\pi}{c}\sqrt{\frac{\mu}{\varepsilon}}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}},\\
E_6 = \frac{2\pi}{c}\sqrt{\frac{\mu}{\varepsilon}}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}}\frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}},\\
H_0 = -\frac{2\pi}{c}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}}\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}, \quad H_2 = -H_4 = -\frac{2\pi}{c}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}},\\
H_6 = \frac{2\pi}{c}\frac{J_{3/2}\Delta z_{3/2}}{1-\beta_{3/2}\mathrm{n}}\frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}}
\end{gathered} \tag{5.66}$$

Перепишем это решение в более компактной форме

$$\begin{aligned}
&E_0 = A_{1/2}\mathcal{E}^J_{3/2}, && E_2 = E_4 = \mathcal{E}^J_{3/2}, && E_6 = A_{5/2}\mathcal{E}^J_{3/2},\\
&H_0 = A_{1/2}\sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^J_{3/2}, && H_2 = -H_4 = \sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^J_{3/2}, && H_6 = -A_{5/2}\sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^J_{3/2}.
\end{aligned} \tag{5.67}$$

Физическая интерпретация эти соотношений аналогична формулам (5.60). Значения поля E_2, H_2 и E_4, H_4 являются граничным условием для волн, распространяющихся соответственно в отрицательном и положительном направлениях оси z. «Набег фазы» этих волн описывается множителями $A_{1/2}$ и $A_{5/2}$. При этом для волн, распространяющихся в отрицательном направлении оси z, выполняется $H = -\sqrt{\varepsilon/\mu}E$. Для волн, распространяющихся в положительном направлении, знак следует изменить на противоположный $H = \sqrt{\varepsilon/\mu}E$.

3° Опираясь на (5.67), нетрудно записать решение для случая, когда сетка содержит произвольное число шагов $N > 3$, но ток отличен от нуля только в шаге n_0: $J_{n_0-1/2} \neq 0$, $J_{n-1/2} = 0$ при $n \neq n_0$. Имеем

$$\begin{aligned}
&E_{2n_0-2-2q} = \prod_{p=1}^{q} A_{n_0-1/2-p}\mathcal{E}^J_{n_0-1/2}, \quad H_{2n_0-2-2q} = -\sqrt{\frac{\varepsilon}{\mu}}\prod_{p=1}^{q} A_{n_0-1/2-p}\mathcal{E}^J_{n_0-1/2},\\
&E_{2n_0-2} = \mathcal{E}^J_{n_0-1/2}, \qquad\qquad H_{2n_0-2} = -\sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^J_{n_0-1/2},\\
&E_{2n_0} = \mathcal{E}^J_{n_0-1/2}, \qquad\qquad H_{2n_0} = \sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^J_{n_0-1/2},\\
&E_{2n_0+2q} = \prod_{p=1}^{q} A_{n_0-1/2+p}\mathcal{E}^J_{n_0-1/2}, \qquad H_{2n_0+2q} = \sqrt{\frac{\varepsilon}{\mu}}\prod_{p=1}^{q} A_{n_0-1/2+p}\mathcal{E}^J_{n_0-1/2}.
\end{aligned} \tag{5.68}$$

Здесь $q \geqslant 1$. Эти формулы следуют из (5.61), (5.62) и (5.64).

4° Если ток присутствует в нескольких ячейках, то нужно просуммировать выражения (5.68) по индексу n_0. При этом вклад в поле E от всех ячеек с током берется со знаком «+». Вклад в поле H в узлах, расположенных левее выбранной ячейки с током, учитывается со знаком «−», а в узлах, расположенных правее выбранной ячейки – со знаком «+».

В качестве примера приведем решение задачи, в которой сетка содержит 4 интервала, причем объемные токи в средних интервалах отличны от нуля $J_{3/2} \neq 0$, $J_{5/2} \neq 0$, а в граничных – равны нулю $J_{1/2} = J_{7/2} = 0$. Это решение имеет вид

$$\begin{aligned}
&E_0 = A_{1/2}\mathcal{E}^J_{1/2} + A_{1/2}A_{3/2}\mathcal{E}^J_{5/2}, \quad H_0 = -\sqrt{\frac{\varepsilon}{\mu}}(A_{1/2}\mathcal{E}^J_{1/2} + A_{1/2}A_{3/2}\mathcal{E}^J_{5/2}),\\
&E_2 = \mathcal{E}^J_{3/2} + A_{3/2}\mathcal{E}^J_{5/2}, \qquad\qquad H_2 = -\sqrt{\frac{\varepsilon}{\mu}}(\mathcal{E}^J_{3/2} + A_{3/2}\mathcal{E}^J_{5/2}),\\
&E_4 = \mathcal{E}^J_{3/2} + \mathcal{E}^J_{5/2}, \qquad\qquad H_4 = \sqrt{\frac{\varepsilon}{\mu}}(\mathcal{E}^J_{3/2} - \mathcal{E}^J_{5/2}),\\
&E_6 = A_{5/2}\mathcal{E}^J_{3/2} + \mathcal{E}^J_{5/2}, \qquad\qquad H_6 = \sqrt{\frac{\varepsilon}{\mu}}(A_{5/2}\mathcal{E}^J_{3/2} + \mathcal{E}^J_{5/2}),\\
&E_8 = A_{7/2}A_{5/2}\mathcal{E}^J_{3/2} + A_{7/2}\mathcal{E}^J_{5/2}, \quad H_8 = \sqrt{\frac{\varepsilon}{\mu}}(A_{7/2}A_{5/2}\mathcal{E}^J_{3/2} + A_{7/2}\mathcal{E}^J_{5/2}).
\end{aligned} \tag{5.69}$$

Оно построено по описанному выше правилу. Можно непосредственно проверить, что это решение удовлетворяет системе

$$\begin{aligned}
&\sqrt{\varepsilon}E_0+\sqrt{\mu}H_0=0, && E_2-E_0-\beta_{1/2}\mu(H_2+H_0)=0,\\
&H_2-H_0-\beta_{1/2}\varepsilon(E_2+E_0)=0, && E_4-E_2-\beta_{3/2}\mu(H_4+H_2)=0,\\
&H_4-H_2-\beta_{3/2}\varepsilon(E_4+E_2)=\frac{4\pi}{c}J_{3/2}\Delta z_{3/2}, && E_6-E_4-\beta_{5/2}\mu(H_6+H_4)=0,\\
&H_6-H_4-\beta_{5/2}\varepsilon(E_6+E_4)=\frac{4\pi}{c}J_{5/2}\Delta z_{5/2}, && E_8-E_6-\beta_{7/2}\mu(H_8+H_6)=0,\\
&H_8-H_6-\beta_{7/2}\varepsilon(E_8+E_6)=0, && \sqrt{\varepsilon}E_8-\sqrt{\mu}H_8=0.
\end{aligned} \tag{5.70}$$

5.5.3. Поверхностные токи

1° Рассмотрим однородную среду $\varepsilon=\text{const}$, $\mu=\text{const}$, $\varepsilon\neq\mu$. Пусть в плоскости $z=z_1=\text{const}$ течет внешний (заданный) поверхностный ток j_1, а индуцированные поверхностные токи равны нулю $\sigma^{\text{surf}}=0$. Пусть сетка содержит $N=2$ интервала, причем точка z_1 является внутренним узлом. Пусть также волны, падающие из бесконечности, отсутствуют $E^0=E^a=0$. Запишем систему разностных уравнений для этой задачи

$$\begin{aligned}
&\sqrt{\varepsilon}E_0+\sqrt{\mu}H_0=0, && E_2-E_0-\beta_{1/2}\mu(H_1+H_0)=0,\\
&H_1-H_0-\beta_{1/2}\varepsilon(E_2+E_0)=0, && H_2-H_1=-\frac{4\pi}{c}j_1,\\
&E_4-E_2-\beta_{3/2}\mu(H_3+H_2)=0, && H_4-H_2-\beta_{3/2}\varepsilon(E_4+E_2)=0,\\
&\sqrt{\varepsilon}E_4-\sqrt{\mu}H_4=0.
\end{aligned} \tag{5.71}$$

Заметим, что левое предельное значение H_1 в узле z_1 отличается от правого предельного значения H_2. Левое и правое предельные значения электрического поля в этом узле совпадают $E_1=E_2$. В узлах z_0, z_2 поля непрерывны.

Решение этой разностной задачи имеет вид

$$\begin{aligned}
&E_0=-\frac{2\pi}{c}j_1\sqrt{\frac{\mu}{\varepsilon}}\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}}, && H_0=\frac{2\pi}{c}j_1\frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}},\\
&E_2=-\frac{2\pi}{c}\sqrt{\frac{\mu}{\varepsilon}}j_1, && H_2=-H_1=-\frac{2\pi}{c}j_1,\\
&E_4=-\frac{2\pi}{c}j_1\sqrt{\frac{\mu}{\varepsilon}}\frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, && H_4=-\frac{2\pi}{c}j_1\frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}.
\end{aligned} \tag{5.72}$$

Это решение подчиняется той же закономерности, что мы видели в предыдущих примерах. Для волны, бегущей в сторону $z = -\infty$, имеем $H = -\sqrt{\varepsilon/\mu}E$. Для волны, бегущей в сторону $z = +\infty$, выполнено $H = \sqrt{\varepsilon/\mu}E$. Поля E_2, H_2 являются граничным условием для волн, распространяющихся в обоих направлениях оси z. При этом смещение от точки расположения источника на один шаг влево приводит к умножению полей на $A_{1/2}$. При смещении вправо поля умножаются на $A_{3/2}$. Логарифмы этих множителей можно трактовать как набеги фаз соответствующих волн. Удобно ввести обозначение

$$\mathcal{E}_1^j = -\frac{2\pi}{c}\sqrt{\frac{\mu}{\varepsilon}}j_1. \tag{5.73}$$

Тогда (5.72) можно записать в более компактной форме

$$\begin{aligned} &E_0 = A_{1/2}\mathcal{E}_1^j, \quad H_0 = -A_{1/2}\sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}_1^j, \\ &E_2 = \mathcal{E}_1^j \qquad\quad H_2 = -H_1 = \sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}_1^j, \\ &E_4 = A_{3/2}\mathcal{E}_1^j \quad H_4 = A_{3/2}\sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}_1^j. \end{aligned} \tag{5.74}$$

2° Рассмотрим теперь задачу на сетке, содержащей $N = 4$ интервала. Пусть поверхностный ток j_2 присутствует в узле z_2, а в прочих узлах равен нулю $j_n = 0$, $n = 0, 1, 3, 4$. Разностная схема имеет вид

$$\begin{aligned} &\sqrt{\varepsilon}E_0 + \sqrt{\mu}H_0 = 0, && E_2 - E_0 - \beta_{1/2}\mu(H_2 + H_0) = 0, \\ &H_2 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = 0, && E_4 - E_2 - \beta_{3/2}\mu(H_3 + H_2) = 0, \\ &H_3 - H_2 - \beta_{3/2}\varepsilon(E_4 + E_2) = 0, && H_4 - H_3 = -\frac{4\pi}{c}j_2, \\ &E_6 - E_4 - \beta_{5/2}\mu(H_6 + H_4) = 0, && H_6 - H_4 - \beta_{5/2}\varepsilon(E_6 + E_4) = 0, \\ &E_8 - E_6 - \beta_{7/2}\mu(H_8 + H_6) = 0, && H_8 - H_6 - \beta_{7/2}\varepsilon(E_8 + E_6) = 0, \\ &\sqrt{\varepsilon}E_8 - \sqrt{\mu}H_8 = 0. \end{aligned} \tag{5.75}$$

Приведем решение этой задачи

$$
\begin{aligned}
E_0 &= -\frac{2\pi}{c} j_2 \sqrt{\frac{\mu}{\varepsilon}} \frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}} \frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, & H_0 &= \frac{2\pi}{c} j_1 \frac{1+\beta_{1/2}\mathrm{n}}{1-\beta_{1/2}\mathrm{n}} \frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, \\
E_2 &= -\frac{2\pi}{c} \sqrt{\frac{\mu}{\varepsilon}} j_2 \frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, & H_2 &= \frac{2\pi}{c} j_2 \frac{1+\beta_{3/2}\mathrm{n}}{1-\beta_{3/2}\mathrm{n}}, \\
E_4 &= -\frac{2\pi}{c} j_1 \sqrt{\frac{\mu}{\varepsilon}}, & H_4 &= -H_3 = -\frac{2\pi}{c} j_1, \\
E_6 &= -\frac{2\pi}{c} \sqrt{\frac{\mu}{\varepsilon}} j_2 \frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}}, & H_6 &= -\frac{2\pi}{c} j_2 \frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}}, \\
E_8 &= -\frac{2\pi}{c} \sqrt{\frac{\mu}{\varepsilon}} j_2 \frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}} \frac{1+\beta_{7/2}\mathrm{n}}{1-\beta_{7/2}\mathrm{n}}, & H_8 &= -\frac{2\pi}{c} j_2 \frac{1+\beta_{5/2}\mathrm{n}}{1-\beta_{5/2}\mathrm{n}} \frac{1+\beta_{7/2}\mathrm{n}}{1-\beta_{7/2}\mathrm{n}},
\end{aligned}
\tag{5.76}
$$

или в более компактной форме

$$
\begin{aligned}
E_0 &= A_{1/2} A_{3/2} \mathcal{E}_2^j, & H_0 &= -A_{1/2} A_{3/2} \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}_2^j, \\
E_2 &= A_{3/2} \mathcal{E}_2^j, & H_2 &= -A_{3/2} \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}_2^j, \\
E_4 &= \mathcal{E}_2^j & H_4 &= -H_3 = \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}_2^j, \\
E_6 &= A_{5/2} \mathcal{E}_2^j & H_6 &= A_{5/2} \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}_2^j, \\
E_8 &= A_{5/2} A_{7/2} \mathcal{E}_2^j & H_8 &= A_{5/2} A_{7/2} \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}_2^j
\end{aligned}
\tag{5.77}
$$

В этом решении видны те же закономерности, что и в предыдущих примерах. Удаление узла на один шаг от точки расположения источника эквивалентно умножению поля на множитель A с индексом, равным номеру соответствующего шага. В точках, для которых координата z больше, чем z-координата источника, электрическое и магнитное поля связаны соотношением $H = \sqrt{\varepsilon/\mu}E$. В точках с z-координатой меньшей, чем z-координата источника, знак последнем соотношении нужно изменить на противоположный $H = -\sqrt{\varepsilon/\mu}E$.

3° Пользуясь решениями (5.74) и (5.61), (5.62), нетрудно составить решение задачи, в которой сетка является произвольной неравномерной и поверхност-

ный ток присутствует только в узле n_0. Оно имеет вид

$$\begin{aligned}
&E_{2n_0-2-2q} = \textstyle\prod_{p=1}^{q} A_{n_0-1/2-p}\mathcal{E}^j_{n_0}, && H_{2n_0-2-2q} = -\sqrt{\frac{\varepsilon}{\mu}}\textstyle\prod_{p=1}^{q} A_{n_0-1/2-p}\mathcal{E}^j_{n_0},\\
&E_{2n_0} = \mathcal{E}^j_{n_0}, && H_{2n_0} = -H_{2n_0-1} = \sqrt{\frac{\varepsilon}{\mu}}\mathcal{E}^j_{n_0},\\
&E_{2n_0+2q} = \textstyle\prod_{p=1}^{q} A_{n_0-1/2+p}\mathcal{E}^j_{n_0}, && H_{2n_0+2q} = \sqrt{\frac{\varepsilon}{\mu}}\textstyle\prod_{p=1}^{q} A_{n_0-1/2+p}\mathcal{E}^j_{n_0}.
\end{aligned} \tag{5.78}$$

Здесь индекс q принимает значения $q \geqslant 1$.

4° Если поверхностные токи текут по нескольким плоскостям $\{z_{n_0}\}$, то необходимо взять сумму по n_0 выражений вида (5.78).

В качестве примера приведем решение задачи, в которой сетка содержит 3 интервала, причем в внутренних узлах $z = z_1$ и $z = z_2$ расположены плоскости с поверхностными токами j_1 и j_2 соответственно. Это решение выглядит следующим образом

$$\begin{aligned}
&E_0 = A_{1/2}\mathcal{E}^j_1 + A_{1/2}A_{3/2}\mathcal{E}^j_2, && E_2 = \mathcal{E}^j_1 + A_{3/2}\mathcal{E}^j_2,\\
&E_4 = A_{3/2}\mathcal{E}^j_1 + \mathcal{E}^j_2, && E_6 = A_{3/2}A_{5/2}\mathcal{E}^j_1 + A_{5/2}\mathcal{E}^j_2,\\
&H_0 = -\sqrt{\frac{\varepsilon}{\mu}}(A_{1/2}\mathcal{E}^j_1 + A_{1/2}A_{3/2}\mathcal{E}^j_2), && H_1 = -\sqrt{\frac{\varepsilon}{\mu}}(\mathcal{E}^j_1 + A_{3/2}\mathcal{E}^j_2),\\
&H_2 = \sqrt{\frac{\varepsilon}{\mu}}(\mathcal{E}^j_1 - A_{3/2}\mathcal{E}^j_2), && H_3 = \sqrt{\frac{\varepsilon}{\mu}}(A_{3/2}\mathcal{E}^j_1 - \mathcal{E}^j_2),\\
&H_4 = \sqrt{\frac{\varepsilon}{\mu}}(A_{3/2}\mathcal{E}^j_1 + \mathcal{E}^j_2), && H_6 = \sqrt{\frac{\varepsilon}{\mu}}(A_{3/2}A_{5/2}\mathcal{E}^j_1 + A_{5/2}\mathcal{E}^j_2).
\end{aligned} \tag{5.79}$$

Прямой подстановкой нетрудно убедиться, что это решение удовлетворяет разностной задаче

$$\begin{aligned}
&\sqrt{\varepsilon}E_0 + \sqrt{\mu}H_0 = 0, && E_2 - E_0 - \beta_{1/2}\mu(H_1 + H_0) = 0,\\
&H_1 - H_0 - \beta_{1/2}\varepsilon(E_2 + E_0) = 0, && H_2 - H_1 = -\frac{4\pi}{c}j_1,\\
&E_4 - E_2 - \beta_{3/2}\mu(H_3 + H_2) = 0, && H_3 - H_2 - \beta_{3/2}\varepsilon(E_4 + E_2) = 0,\\
&H_4 - H_3 = -\frac{4\pi}{c}j_2, && E_6 - E_4 - \beta_{5/2}\mu(H_6 + H_4) = 0,\\
&H_6 - H_4 - \beta_{5/2}\varepsilon(E_6 + E_4) = 0, && \sqrt{\varepsilon}E_6 - \sqrt{\mu}H_6 = 0.
\end{aligned} \tag{5.80}$$

5.5.4. Суперпозиция решений

В общем случае поле может возбуждаться волнами $E^0 \neq 0$, $E^a \neq 0$, приходящими из $z = -\infty$ и $z = +\infty$, объемными $\{J_{n_0-1/2}\}$ и поверхностными $\{j_{n_0}\}$ токами. Тогда точное решение системы разностных уравнений представляется суммой выражений (5.61), (5.62), (5.68), (5.78), причем две последних формулы суммируются по индексам n_0, соответствующим ненулевым токам.

Построенное решение является новым. Оно записывается по явным формулам (в указанных выше формулах поля выражены друг через друга только для компактности записи). Это решение применимо только для однородной среды $\varepsilon = \text{const}$, $\mu = \text{const}$. При этом сетка $\{\Delta z_{n-1/2}\}$ может быть произвольной неравномерной. Конфигурация объемных и поверхностных токов также может быть произвольной.

5.6. Метод спектрального разложения

5.6.1. Формулировка метода

1° Как известно, при распространении волнового пакета в линейной диспергирующей среде для разных спектральных компонент решения реализуются разные значения ε и μ и, соответственно, разные скорости распространения. В [264] приведено решение задачи о распространении электромагнитного импульса в линейной диспергирующей среде. Оно состоит из трех этапов: фурье-анализа исходного сигнала, преобразования каждой спектральной компоненты поля диспергирующей средой, фурье-синтеза сигнала на выходе.

2° В данной работе предлагается следующий подход. Применим описанный выше алгоритм к нестационарной задаче (5.25) – (5.29). Разложим падающие волновые пакеты, а также объемные и поверхностные токи на монохроматиче-

ские компоненты с помощью прямого преобразования Фурье

$$\begin{aligned}
\tilde{f}^{0,a}(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f^{0,a}(\zeta) e^{-i\omega\zeta} d\zeta, \\
\tilde{J}_x^{\text{ext}}(\omega, z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} J_x^{\text{ext}}(t, z) e^{-i\omega t} dt, \\
\tilde{j}_x^{\text{ext}}(\omega, z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} j_x^{\text{ext}}(t, z) e^{-i\omega t} dt.
\end{aligned} \tag{5.81}$$

Решим стационарную задачу (5.21) – (5.24) для каждой частоты ω. Полученные решения обозначим $\tilde{E}_x(\omega, z)$, $\tilde{H}_y(\omega, z)$. Затем выполним обратное преобразование Фурье

$$\begin{aligned}
E_x(t, z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \tilde{E}_x(\omega, z) e^{i\omega t} d\omega, \\
H_y(t, z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \tilde{H}_y(\omega, z) e^{i\omega t} d\omega.
\end{aligned} \tag{5.82}$$

Данный подход является обобщением теории, изложенной в [264].

3° Численная реализация этого алгоритма имеет следующий вид. Выполним численное преобразование Фурье (5.81) падающих волновых пакетов, объемных и поверхностных токов по формуле трапеций, используя одинаковые сетки по времени и наборы частот $\{\omega_m\}$, $\omega_{m+1} - \omega_m = \Delta\omega_m$

$$\begin{aligned}
\tilde{f}^{0,a}(\omega_m) &= \sum_{k=0}^{K} f^{0,a}(\zeta_k) e^{-i\omega t_k} g_k \Delta\zeta_k, \\
\tilde{J}_x^{\text{ext}}(\omega_m, z_n) &= \sum_{k=0}^{K} J_x^{\text{ext}}(t_k, z_n) e^{-i\omega t_k} g_k \Delta t_k, \\
\tilde{j}_x^{\text{ext}}(\omega_m, z_n) &= \sum_{k=0}^{K} j_x^{\text{ext}}(t_k, z_n) e^{-i\omega t_k} g_k \Delta t_k.
\end{aligned} \tag{5.83}$$

Здесь $g_0 = g_K = 0.5$, $g_k = 1$, $k \neq 1, K$ – веса квадратурной формулы трапеций.

По описанному выше методу решим набор стационарных задач (5.21) – (5.24), соответствующих различным частотам ω_m. Соответствующие решения обозначим $\tilde{E}(\omega_m, z_n)$, $\tilde{H}(\omega_m, z_n)$. Напомним, что в каждой из этих задач реализуются свои значения $\varepsilon(\omega_m, z_n)$, $\mu(\omega_m, z_n)$, $\sigma(\omega_m, z_n)$. Поэтому предлагаемый подход позволяет учитывать произвольный гладкий закон частотной дисперсии.

Просуммируем полученные спектральные амплитуды решения по всем частотам ω_m

$$\begin{aligned} E(t_k, z_n) &= \textstyle\sum_{m=0}^{M} \tilde{E}(\omega_m, z_n) g_m \Delta\omega_m, \\ H(t_k, z_n) &= \textstyle\sum_{m=0}^{M} \tilde{H}(\omega_m, z_n) g_m \Delta\omega_m. \end{aligned} \tag{5.84}$$

Это дает решение $E(z,t)$, $H(z,t)$ исходной нестационарной задачи.

Назовем этот алгоритм ***методом спектрального разложения***.

5.6.2. Сходимость

1° Если функции $E^{0,a}$, $j_q(z,t)$, $J_q(z,t)$ имеют вторые непрерывные производные, то квадратуры, выражающие прямое и обратное преобразование Фурье, сходятся со вторым порядком точности: $O(\Delta\omega^2)$ и $O(\Delta t^2)$ соответственно. Как показано выше, аппроксимация разностных схем для стационарных задач есть $O(\Delta z^2)$. Поэтому аппроксимация схемы есть $O(\Delta z^2 + \Delta\omega^2 + \Delta t^2)$.

2° Покажем, что метод спектрального разложения устойчив относительно малых возмущений входных данных. Предварительно заметим, что в реальных задачах падающие импульсы и токи являются локализованными функциями времени: отрезок T, на котором они существенно отличны от нуля, является конечным. Соответственно, интервал частот Ω, на котором отличны от нуля их фурье-образы, также конечен. Чаще всего используется временная модуляция гауссовой огибающей. Хотя она формально отлична от нуля на всей прямой $-\infty < t < +\infty$, однако такая огибающая очень быстро убывает по мере удаления от максимума. Для получения физически разумной точности 0.1% достаточно взять $T \sim 3\sigma$, где σ – ширина огибающей на половине высоты. Машинная точность 10^{-16} достигается при $T \sim 6\sigma$.

Пусть в одну из функций $f^{0,a}$, j_q, J_q внесена ошибка δ, зависящая от t и, возможно, z. Погрешность соответствующего фурье-образа равна

$$\tilde{\delta}(\omega, z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \delta(t, z) e^{-i\omega t} dt \approx \sum_{k=0}^{K} \delta(t_k, z) e^{-i\omega t_k} g_k \Delta t_k. \tag{5.85}$$

Легко видеть, что погрешность (5.85) не превосходит

$$|\tilde{\delta}| \le \frac{1}{\sqrt{2\pi}} \max |\delta| T \tag{5.86}$$

Такое возмущение входных данных приводит к возмущению $\delta\tilde{E}$, $\delta\tilde{H}$ решений стационарных задач, для которого справедливы оценки (5.47), (5.50), (5.52), в которые вместо исходных входных данных следует подставить величину (5.86).

Выполним обратное преобразование Фурье погрешностей $\delta\tilde{E}$, $\delta\tilde{H}$

$$\delta E(t,z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \delta\tilde{E}(\omega,z) e^{i\omega t} dt \approx \sum_{m=0}^{M} \delta\tilde{E}(\omega_m,z) g_m \Delta\omega_m. \tag{5.87}$$

Справедлива следующая оценка:

$$|\delta E| \le \frac{1}{2\pi} C_{\max} \max |\delta| T\Omega. \tag{5.88}$$

Здесь $C_{\max}$ – наибольшая из мажорирующих констант в оценках (5.47), (5.50), (5.52). Аналогично выписывается оценка для погрешности δH. Это завершает обоснование устойчивости предложенной схемы. Оно проведено для произвольных неравномерных сеток по всем переменным и неоднородных сред (в том числе, диспергирующих).

3° Схема является однозначно разрешимой при произвольных входных данных (поскольку однозначно разрешимы разностные схемы для всех стационарных подзадач). Следовательно, она корректна. Поэтому, согласно классическим теоремам теории разностных схем, справедливо следующее утверждение:

Теорема 18. *Схема метода спектрального разложения сходится со вторым порядком точности.* •

4° Для предложенного метода применимы оценки точности по правилу Ричардсона и экстраполяционное уточнение. Заметим, что сетки по переменным z, t, ω необходимо сгущать одновременно и в одинаковое число раз.

Несмотря на простоту, метод спектрального разложения является принципиально новым. Он позволяет учитывать произвольный закон дисперсии, включая таблично заданный. Этот метод рекомендуется использовать как общий

подход построения разностных схем для гиперболических задач в линейной диспергирующей среде.

5.7. Основные результаты главы

1. Для системы стационарных одномерных уравнений Максвелла построена бикомпактная разностная схема, шаблон которой включает только один шаг пространственной сетки. Границы слоев выбираются в качестве узлов сетки, поэтому схема позволяет проводить расчеты обобщенных решений с разрывом решения и его производной. Консервативность схемы обеспечивает сходимость к правильному обобщенному решению.
2. Для ряда важных частных случаев построено явное решение системы разностных уравнений бикомпактной схемы в конечном виде. Это решение является новым. Оно применимо для однородной среды, произвольной неравномерной сетки и произвольной конфигурации объемных и поверхностных токов.
3. Для нестационарных задач предложен принципиально новый метод спектрального разложения, позволяющий учитывать частотную дисперсию материала. Этот подход имеет простую физическую интерпретацию. Он применим для других линейных гиперболических задач в диспергирующих средах.
4. Для предложенных методов строго доказана сходимость для произвольных неравномерных сеток и неоднородных сред.

6. Верификация бикомпактных схем для одномерных уравнений Максвелла

6.1. Стационарные задачи

Для непроводящих (диэлектрических) сред удалось построить ряд тестовых с задач, у которых точное решение выражается в конечном виде через элементарные функции. В приведенных ниже модельных задачах все размерные величины нормированы (т.е. являются относительными).

6.1.1. Однородная среда

1° Рассмотрим простейшую задачу – распространение плоской монохроматической волны через однородную диэлектрическую среду. Пусть ε, μ постоянны, и объемные и поверхностные токи отсутствуют $J = 0$, $j = 0$. Пусть слева на выделенную область $0 \leqslant z \leqslant a$ падает плоская монохроматическая волна.

2° Решение удобно строить с помощью перехода к волновому уравнению. Запишем одномерные уравнения Максвелла в дифференциальной форме

$$\frac{\partial E_x}{\partial z} = \frac{i\mu\omega}{c} H_y; \quad \frac{\partial H_y}{\partial z} = \frac{i\varepsilon\omega}{c} E_z, \quad 0 \leqslant z \leqslant a; \tag{6.1}$$

$$\frac{\partial E_x}{\partial z} + i\tilde{k} E_x = 2i\tilde{k} E^0, z = 0; \quad \frac{\partial E_x}{\partial z} - i\tilde{k} E_x = 0, z = a. \tag{6.2}$$

Здесь $\tilde{k} = \omega\sqrt{\varepsilon\mu}/c$ – волновое число в среде. В записи граничных условий используется именно это значение, поскольку за пределами отрезка $[0, a]$ волна распространяется не в вакууме, а в той же самой среде.

Исключим из (6.1) H_y. Для этого продифференцируем первое из уравнений по z, выразим из него $\partial H_y/\partial z$ и подставим во второе уравнение. Получим

$$\frac{\partial^2 E_x}{\partial z^2} + \tilde{k}^2 E_x = 0. \tag{6.3}$$

Общее решение (6.3) есть

$$E_x = C_1 \exp(i\tilde{k}z) + C_2 \exp(-i\tilde{k}z). \tag{6.4}$$

С учетом граничных условий (6.2) имеем $C_1 = E^0$, $C_2 = 0$. Решение H_y найдем с помощью первого из уравнений (6.1). Таким образом, окончательное решение имеет вид

$$E_x = E^0 e^{i\tilde{k}z}, \quad H_y = E^0 \sqrt{\frac{\varepsilon}{\mu}} e^{i\tilde{k}z}. \tag{6.5}$$

Оно известно из учебников.

3° Приведем результаты численных расчетов по предложенной стационарной схеме. Положим $E^0 = 1$, $\omega = \pi/2$, $c = 1$, $a = 1$. Рассмотрим сначала среду с поглощением: $\varepsilon = 1 + 5i$, $\mu = 5$. Такую задачу можно трактовать как жесткую, поскольку решение имеет значительную пространственную разномасштабность (см. п. 1.1.3). На отрезке $0 < z < 0.2$ оно меняется в e раз. Решение этой задачи представлено на рис. 6.1.

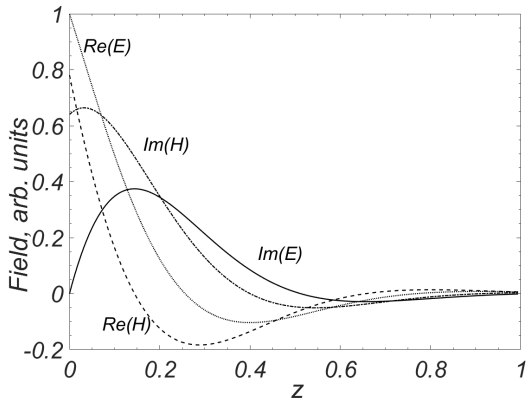

Рис. 6.1. Решение задачи о среде с поглощением.

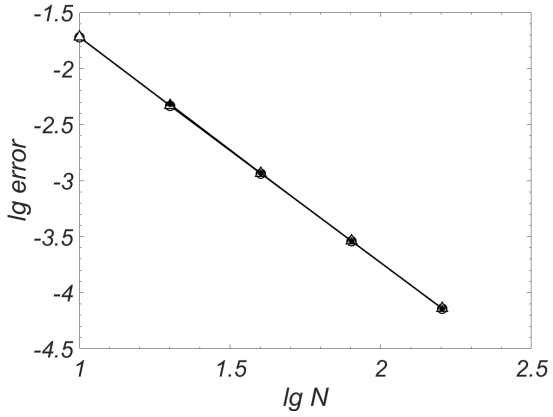

Рис. 6.2. Погрешность решения в задаче о среде с поглощением: $\circ$ – E_x, $\triangle$ – H_y, светлые маркеры – разность численного и точного решений, темные маркеры – оценки по методу Ричардсона.

Расчеты проводились на наборе сгущающихся сеток. Погрешность расчета определялась двумя способами: во-первых, как разность численного решения и точного, во-вторых, с помощью апостериорной оценки по методу Ричардсона. Нормы L_2 полученных относительных погрешностей для полей E_x и H_y приведены на рис. 6.2 в двойном логарифмическом масштабе. Видно, что для

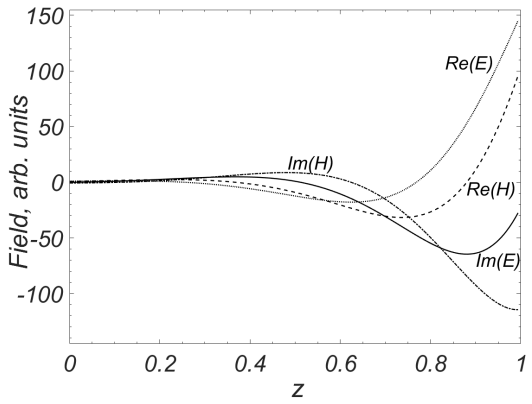

Рис. 6.3. Решение задачи об активной среде.

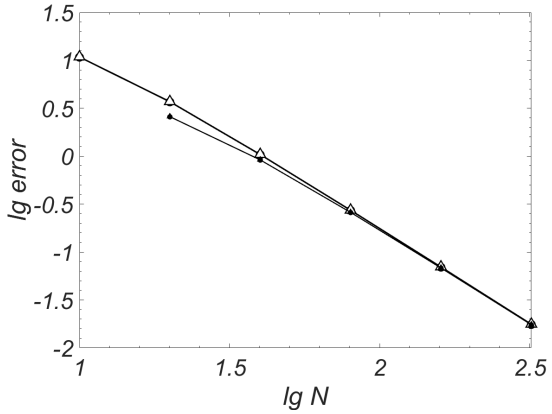

Рис. 6.4. Погрешность решения в задаче об активной среде. Обозначения соответствуют рис. 6.2.

обоих полей погрешность убывает. Линии погрешности являются прямыми с наклоном -2. Поэтому фактический порядок точности равен 2, что совпадает с теорией. Видно также, что оценка точности по методу Ричардсона практически совпадает с погрешностью, равной разности численного решения и точного.

4° Аналогичные расчеты проводились для активной среды: $\varepsilon = 1 - 5i$, $\mu = 5$. В этом случае амплитуды полей при $z = 1$ оказываются в ≈ 150 раз больше, чем при $z = 0$. Такая задача является плохо обусловленной по левому граничному условию. Вид решения показан на рис. 6.3, а полученные относительные погрешности – на рис. 6.4. В этом случае справедливы те же выводы, что и в предыдущем. Начальные участки графиков погрешности искривлены, но по мере сгущения сеток погрешность начинает убывать в соответствии с теоретическим вторым порядком точности.

6.1.2. Граница раздела диэлектриков

1° Пусть расчетная область состоит из двух однородных недиспергирующих сред с границей в точке $z = b$. Пусть слева от нее материальные параметры равны ε_1, μ_1, а справа – ε_2, μ_2. Пусть волна падает слева $E^0 = 1$, $E^a = 0$, а токи отсутствуют $J_{1,2} = 0$, $j_1 = 0$. Постановка задачи в форме одномерных

дифференциальных уравнений Максвелла имеет следующий вид:

$$\frac{\partial E_x}{\partial z} = \frac{i\mu_1\omega}{c} H_y; \quad \frac{\partial H_y}{\partial z} = \frac{i\varepsilon_1\omega}{c} E_x, \quad 0 \leqslant z \leqslant b; \tag{6.6}$$

$$\frac{\partial E_x}{\partial z} = \frac{i\mu_2\omega}{c} H_y; \quad \frac{\partial H_y}{\partial z} = \frac{i\varepsilon_2\omega}{c} E_x, \quad b \leqslant z \leqslant a; \tag{6.7}$$

$$\frac{\partial E_x}{\partial z} + \frac{i\omega}{c} E_x = 2i\tilde{k}E^0, z = 0; \quad \frac{\partial E_x}{\partial z} - \frac{i\omega}{c} E_x = 0, z = a. \tag{6.8}$$

$$E_x|_{z=b-0} = E_x|_{z=b+0}, \quad H_y|_{z=b-0} = H_y|_{z=b+0}. \tag{6.9}$$

2° Получим точное решение этой задачи. От (6.6) и (6.7) перейдем к уравнениям типа (6.3) для областей $0 \leqslant z \leqslant b$ и $b \leqslant z \leqslant a$. Общее решение имеет вид

$$E_x = C_1 e^{i\tilde{k}_1 z} + C_2 e^{-i\tilde{k}_1 z}, \; H_y = \sqrt{\frac{\varepsilon_1}{\mu_1}}(C_1 e^{i\tilde{k}_1 z} - C_2 e^{-i\tilde{k}_1 z}), \; 0 \leqslant z \leqslant b; \tag{6.10}$$

$$E_x = C_3 e^{i\tilde{k}_2 z} + C_4 e^{-i\tilde{k}_2 z}, \; H_y = \sqrt{\frac{\varepsilon_2}{\mu_2}}(C_3 e^{i\tilde{k}_2 z} - C_4 e^{-i\tilde{k}_2 z}), \; b \leqslant z \leqslant a. \tag{6.11}$$

Из граничных условий (6.8) следует, что

$$C_1 = E^0 = 1, \quad C_4 = 0. \tag{6.12}$$

Подстановка в условия сопряжения дает

$$C_2 = E^0 \frac{\sqrt{\varepsilon_1/\mu_1} - \sqrt{\varepsilon_2/\mu_2}}{\sqrt{\varepsilon_1/\mu_1} + \sqrt{\varepsilon_2/\mu_2}} e^{i\tilde{k}_1 b}, \quad C_3 = E^0 \frac{2\sqrt{\varepsilon_1/\mu_1}}{\sqrt{\varepsilon_1/\mu_1} + \sqrt{\varepsilon_2/\mu_2}} e^{i(\tilde{k}_1 - \tilde{k}_2)b}. \tag{6.13}$$

Выражения (6.13) являются частным случаем общеизвестных формул Френеля.

3° Положим $c = 1, \omega = \pi, a = 1, b = 0.5$. Пусть для левой среды $\varepsilon_1 = \mu_1 = 1$, а для правой – $\varepsilon_2 = 10$, $\mu_2 = 2$. На рис. 6.5 представлен вид решения. Видно, что при $z = b$ возникает слабый разрыв (излом) у $\operatorname{Im} E$ и $\operatorname{Im} H$, а $\operatorname{Re} H$ и $\operatorname{Re} H$ являются непрерывными и гладкими.

Приведем результаты расчета по стационарной бикомпактной схеме. Выберем специальную сетку, у которой узел попадает в точку $z = b$. Очевидно, все последующие сетки, полученные последовательным уменьшением шага вдвое,

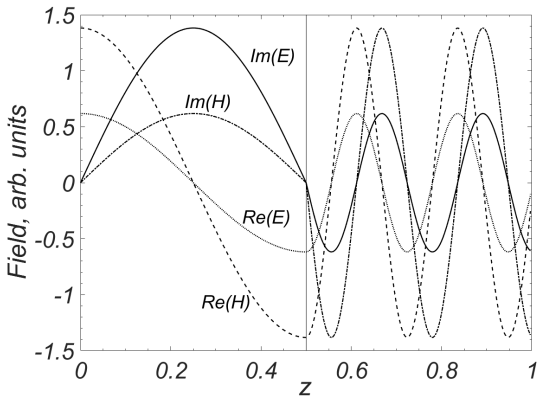

Рис. 6.5. Решение задачи о диэлектрической границе раздела. Вертикальная прямая – граница раздела.

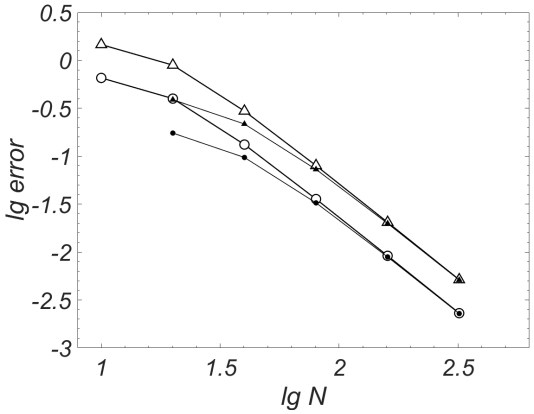

Рис. 6.6. Погрешность решения в задаче о диэлектрической границе раздела. Обозначения соответствуют рис. 6.2.

также являются специальными. На рис. 6.6 показана зависимость относительной погрешности в норме L_2 от числа шагов сетки. Масштаб графика двойной логарифмический. Видно, что погрешности, полученная сравнением численного решения с точным, практически совпадают с оценками по методу Ричардсона. Эти кривые стремятся к прямым линиям с наклоном -2, соответствующим теоретическому порядку точности. Напомним, что именно это обеспечивает применимость метода Ричардсона.

4° Сравним предлагаемую схему с методом конечных элементов на примере пакета FreeFEM++ [306]. Расчеты проводились с использованием следующих элементов:

- лагранжевы элементы L1 1-го порядка,
- пузырьковые (B, bubble) элементы 1-го порядка [307],
- разрывные (NC, non-conforming) элементы 1-го порядка,
- лагранжевы элементы L2 2-го порядка.

Для простоты мы задавали граничные условия для обоих полей E, H по значению точного решения на левой границе

$$E_x|_{z=0} = E_x^{\text{ex}}(0), \quad H_y|_{z=0} = H_y^{ex}(0). \tag{6.14}$$

Такая постановка соответствует, например, [308, 309].

Расчет проводился со сгущением сетки, причем каждая из сеток была специальной (т.е. использовались те же сетки, что в расчетах по бикомпактной схеме). На каждой из них вычислялись погрешности δE, δH относительно точного решения. Чтобы не загромождать график, мы проводили дополнительное усреднение погрешностей по компонентам полей

$$\delta = \frac{1}{2}(\delta E + \delta H). \tag{6.15}$$

Величины (6.15) для перечисленных видов конечных элементов приведены на рис. 6.7. Масштаб графика двойной логарифмический. Видно, что элементы L2 обеспечивают 2-ой порядок точности, причем кривая погрешности является плавной. Для элементов L1 порядок точности в среднем близок ко 2-му, но кривая оказывается дерганой, то есть отсутствует на ней прямолинейный участок теоретической сходимости. Это затрудняет контроль точности по Ричардсону и ухудшает надежность расчета. Элементы B и NC обеспечивают только первый порядок точности и значительно уступают по точности элементам L1 и L2.

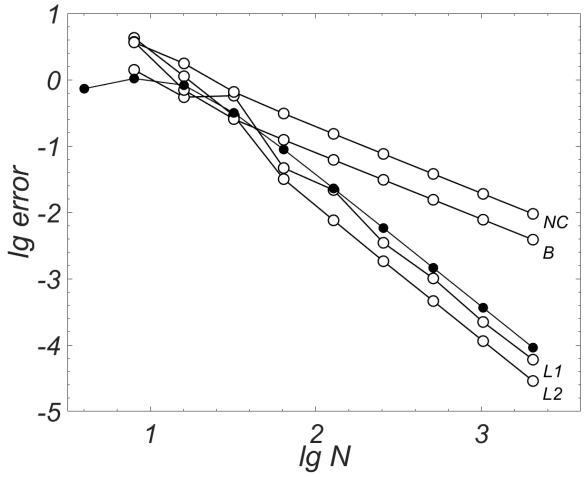

Рис. 6.7. Усредненные погрешности (6.15) в задаче о диэлектрической границе раздела. • – бикомпактная схема, ∘ – метод конечных элементов в пакете FreeFEM++. Типы элементов указаны у кривых.

На рис. 6.7 приведена также усредненная погрешность бикомпактной схемы. Видно, что эта схема дает несколько худшую точность, чем элементы L2, и практически не уступает в точности элементам L1. При этом трудоемкость

бикомпактной схемы такова же, как у элементов L1, и существенно меньше, чем у элементов L2.

5° Таким образом, предложенная схема уверенно справляется с задачами, в которых решение испытывает слабый разрыв. При этом она практически не уступает методу конечных элементов.

6.1.3. Граница раздела с токами

1° Пусть в предыдущей задаче отсутствуют падающие волны $E^0 = E^a = 0$ и объемные токи $J_{1,2} = 0$, но по границе раздела течет поверхностный ток, излучающий волны в положительном и отрицательном направлении оси z. Такая модель описывает переизлучение тонкого металлического напыления на границе рассеивателя. Эта задача особенно трудна, поскольку у $\mathrm{Re}\, H$ возникает сильный разрыв, а у $\mathrm{Im}\, E$ и $\mathrm{Im}\, H$ – слабый. В литературе такая задача не рассматривалась.

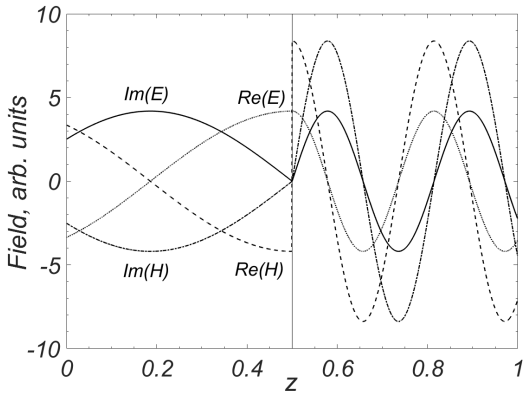

Рис. 6.8. Решение задачи о границе раздела с поверхностным током. Вертикальная прямая – граница раздела.

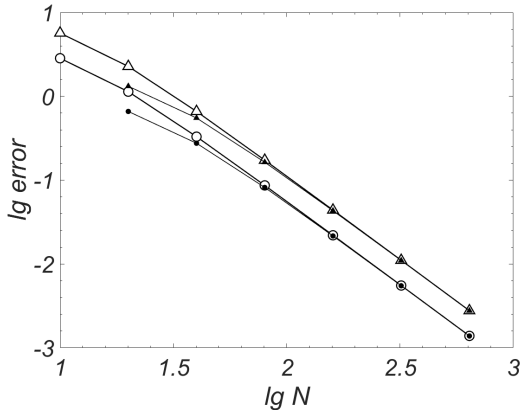

Рис. 6.9. Погрешность решения в задаче о границе раздела с поверхностным током. Обозначения соответствуют рис. 6.2.

Постановка задачи включает уравнения (6.6), (6.7), граничные условия (6.8); условия сопряжения на границе раздела $z = b$ формулируются следующим образом:

$$E_x|_{z=b-0} = E_x|_{z=b+0}, \quad H_y|_{z=b+0} - H_y|_{z=b-0} = \frac{4\pi}{c} j_1. \tag{6.16}$$

2° Выведем точное решение этой задачи. Перейдем от уравнений Максвелла к уравнениям типа (6.3) для областей $0 \leqslant z \leqslant b$ и $b \leqslant z \leqslant a$. Общее решение этих уравнений имеет вид (6.10), (6.11). Волны, приходящие из бесконечности, отсутствуют, поэтому граничные условия дают

$$C_1 = C_4 = 0. \tag{6.17}$$

Подставим эти решения в условия сопряжения (6.16) и найдем C_2, C_3

$$C_2 = \frac{4\pi}{c} j_1 \frac{e^{-i\tilde{k}_1 b}}{\sqrt{\varepsilon_1/\mu_1} + \sqrt{\varepsilon_2/\mu_2}}, \quad C_3 = \frac{4\pi}{c} j_1 \frac{e^{-i\tilde{k}_2 b}}{\sqrt{\varepsilon_1/\mu_1} + \sqrt{\varepsilon_2/\mu_2}}. \tag{6.18}$$

Полученное решение приведено на рис. 6.8. Оно является новым.

3° В расчетах положим $j = 1$, а все остальные параметры выберем такими же, как в предыдущей задаче. Выберем специальную сетку указанным выше способом. На рис. 6.9 показана зависимость погрешности в норме L_2 от числа шагов сетки в двойном логарифмическом масштабе. Видно, что значения погрешности, вычисленные по методу Ричардсона, практически совпадают с погрешностями относительно точного решения. Скорость убывания погрешности соответствует теоретическому второму порядку точности.

4° Был проведен расчет этой задачи в пакете FreeFEM++. Граничные условия задавались аналогично (6.14). Использовались те же 4 типа конечных элементов.

В пакете FreeFEM++ не предусмотрено задание точечных источников, поэтому для имитации поверхностного тока вводился объемный ток следующего вида

$$J = \frac{2\pi}{c}\delta(z - a/2) \approx \frac{4\pi}{cd} \begin{cases} \dfrac{x - x_l}{a/2 - x_l}, & x_l < x < a/2; \\ \dfrac{x - l/2}{x_r - a/2}, & a/2 < x < x_r; \\ 0, & 0 < x < x_l,\ x_r < x < a. \end{cases} \tag{6.19}$$

Здесь $d \ll a$, $x_l = a/2 - d/2$, $x_r = a/2 + d/2$. Ток (6.19) является кусочно-линейной функцией координаты. Он отличен от нуля вне отрезка $[x_l, x_r]$, нарас-

тает на отрезке $[x_l, a/2]$ и убывает на отрезке $[a/2, x_r]$. При этом выполняется условие нормировки

$$\int_0^a J dz = \frac{2\pi}{c}. \tag{6.20}$$

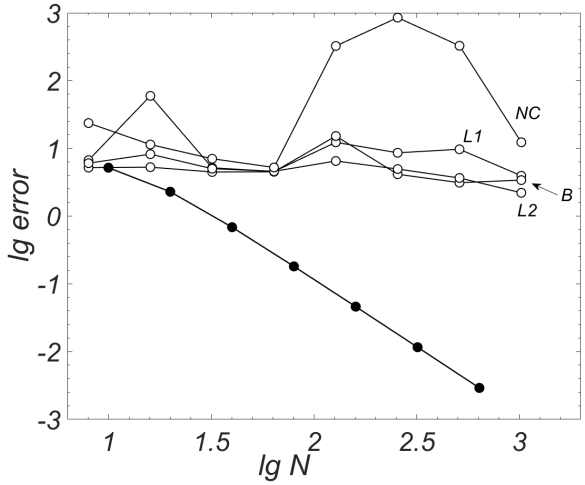

Рис. 6.10. Усредненные погрешности (6.15) в задаче в задаче о границе раздела с поверхностным током. Обозначения соответствуют рис. 6.7.

Погрешности, полученные в пакете FreeFEM++ в данной задаче, приведены на рис. 6.10 в двойном логарифмическом масштабе. Здесь выбрано $\delta = 10^{-3}$. Видно, что для всех рассмотренных типов конечных элементов погрешность велика: она составляет $\sim 10^1$. При сгущении сеток погрешность не убывает, то есть сходимость отсутствует. Аналогичные результаты имели место и при других значениях $d = 10^{-1}$, 10^{-2}, 10^{-5}.

На рис. 6.10 приведена также усредненная погрешность бикомпактной схемы. Видно, что она убывает в соответствии с теоретическим 2-м порядком сходимости. Точность бикомпактной схемы кардинально превосходит пакет FreeFEM++.

5° Таким образом, предлагаемая бикомпактная схема на специальных сетках позволяет проводить расчеты решений с сильными разрывами. Это свидетельствует об исключительно высокой надежности этого метода. Пакет FreeFEM++, основанный на методе конечных элементов, не позволил решить эту задачу. Это убедительно показывает преимущества предлагаемой схемы.

6.1.4. Плотность энергии электромагнитного поля

Рассмотрим плотность энергии электромагнитного поля

$$\mathcal{E} = \frac{1}{8\pi}(\varepsilon|E_x|^2 + \mu|H_y|^2). \tag{6.21}$$

Если среда прозрачная (то есть поглощение отсутствует), то энергия сохраняется при распространении волны.

В задачах п. 6.1.2, 6.1.3 мы вычисляли $\mathcal{E}$ в каждом узле сетки. Затем вычислялась L_2-норма разности численного $\mathcal{E}_n$ и точного $\mathcal{E}^{\text{ex}}$ значений этой величины

$$\delta\mathcal{E} = \|\mathcal{E}^{\text{ex}}(z_n) - \mathcal{E}_n\|_{L_2}. \tag{6.22}$$

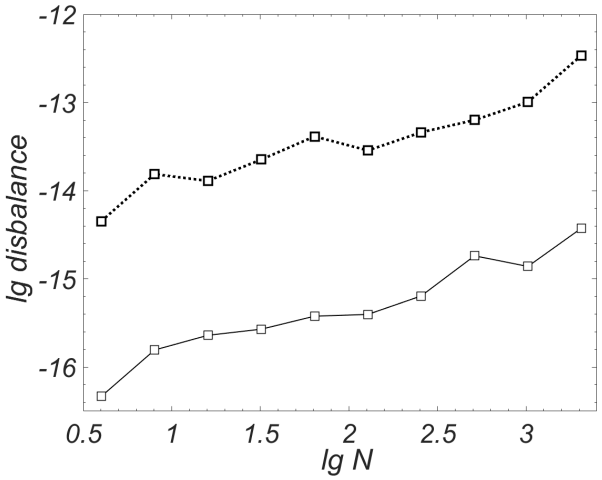

Рис. 6.11. Дисбаланс энергии (6.22) для бикомпактной схемы. Сплошная линия – задача 6.1.2, пунктир – задача 6.1.3.

Величину $\delta\mathcal{E}$ можно интерпретировать как средний дисбаланс энергии электромагнитного поля. Расчеты велись со сгущением сеток, и средний дисбаланс вычислялся на каждой сетке. На рис. 6.11 приведена зависимость $\delta\mathcal{E}$ от числа узлов сетки для обеих задач 6.1.2, 6.1.3. Масштаб графика двойной логарифмический. Видно, что в задаче 6.1.2 дисбаланс составляет $\delta\mathcal{E} \sim 3\cdot10^{-17} \div 3\cdot10^{-15}$, а в задаче 6.1.3 – $\delta\mathcal{E} \sim 3\cdot10^{-15} \div 3\cdot10^{-13}$. С увеличением N дисбаланс возрастает, причем средний наклон кривых равен $1/2$. Это соответствует закону возрастания $\delta\mathcal{E} \sim N^{1/2}$. Такие значения дисбаланса и характер его роста соответствуют ошибкам компьютерного округления.

Таким образом, независимо от величины шага сетки предложенная бикомпактная схема передает закон сохранения энергии электромагнитной волны с точностью ошибок округления. Это является дополнительным преимуществом данной схемы.

6.1.5. Фотонный кристалл

Постановка задачи. 1° Наиболее представительной проверкой любого численного метода является расчет реальной задачи и сравнение с результатами эксперимента. Примером одномерной задачи является распространение излучения через одномерный фотонный кристалл (ФК) при нормальном падении.

Рассмотрим ФК, состоящий из 7 пар слоев {SiO_2 – 160 нм, Ta_2O_5 – 112 нм}. Поверх последнего слоя Ta_2O_5 нанесен слой SiO_2 толщиной 260 нм (см. рис. 6.12) [310]. ФК находится в воздухе. На него из $z = -\infty$ нормально падает плоская линейно поляризованная монохроматическая волна.

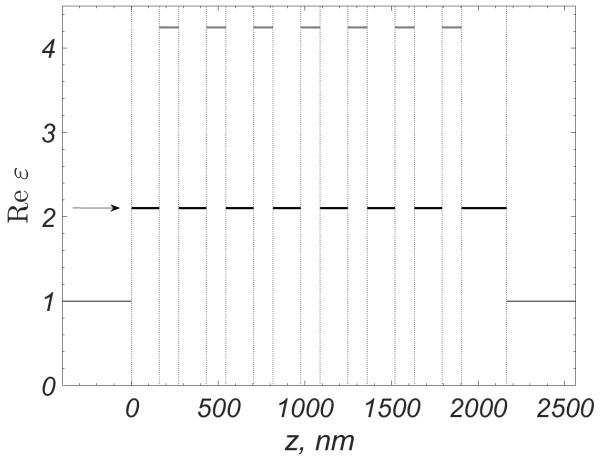

Рис. 6.12. Фотонный кристалл. Зависимость ε от координаты z, соответствующая длины волны на длине волны $\lambda = 900$ нм. Вертикальные линии – границы слоев. Стрелка – направление распространения падающей волны.

2° Закон дисперсии указанных материалов (то есть зависимость материальных параметров от длины волны) были взяты из широко известной библиотеки [311]. Оригинальные измерения выполнены в работе [312]. На рис. 6.13 приведен закон дисперсии в виде зависимости диэлектрической проницаемо-

сти от длины волны. На рис. 6.14 представлен этот же закон дисперсии в виде зависимости показателя преломления от длины волны.

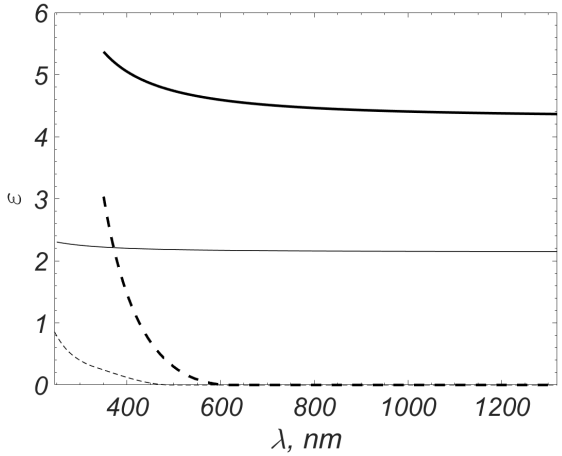

Рис. 6.13. Зависимость диэлектрической проницаемости ε от длины волны λ [311, 312]. Жирные линии – Ta_2O_5, тонкие – SiO_2. Сплошные линии – $\mathrm{Re}\,\varepsilon$, штриховые – $10^3 \cdot \mathrm{Im}\,\varepsilon$.

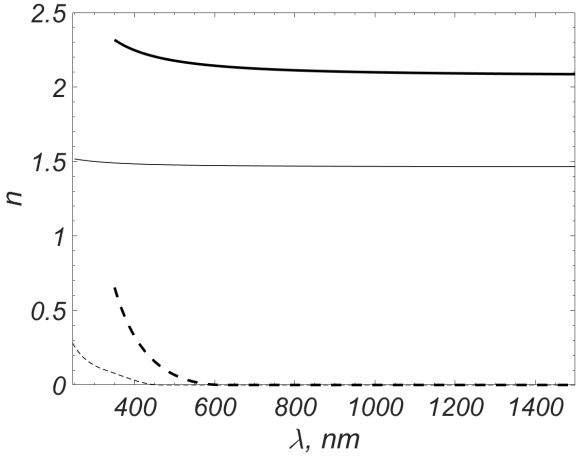

Рис. 6.14. Зависимость показателя преломления n от длины волны λ [311, 312]. Жирные линии – Ta_2O_5, тонкие – SiO_2. Сплошные линии – $\mathrm{Re}\,\mathrm{n}$, штриховые – $10^3 \cdot \mathrm{Im}\,\mathrm{n}$

Расчетные спектры. 1° Были вычислены спектры отражения и прохождения этого ФК. Они приведены на рис. 6.15 и 6.16.

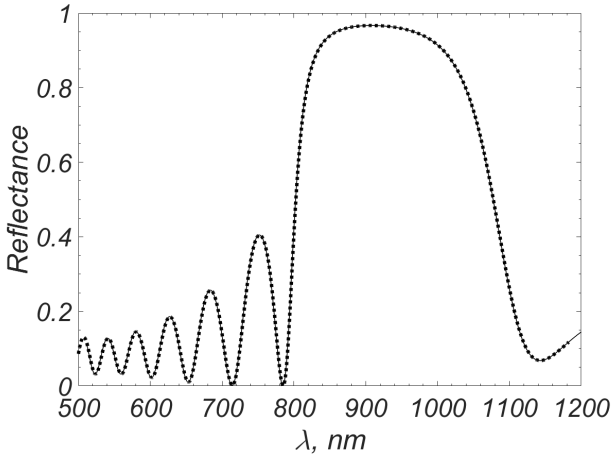

Рис. 6.15. Расчетный спектр отражения ФК на рис. 6.12. Сплошная линия – бикомпактная схема, пунктир – матричный метод [313, 314].

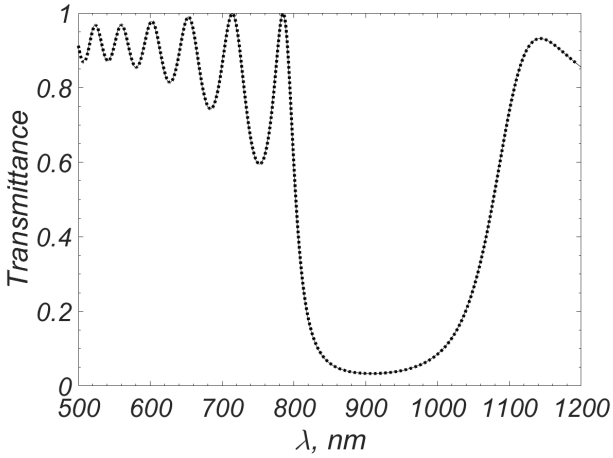

Рис. 6.16. Расчетный спектр прохождения ФК на рис. 6.12. Обозначения соответствуют рис. 6.15.

Опишем процедуру расчета. Зададим диапазон длин волн $500 \leqslant \lambda \leqslant 1200$ нм и введем равномерную сетку по длине волны. Решим задачу при каждом сеточном λ, полагая амплитуду падающей волны равной единице. Найдем отраженную волну E_{refl} как разность численного решения левее крайней левой пластины и падающей волны. Вычислим интенсивность отраженной волны $|E_{\text{refl}}|^2$. Эта величина должна быть постоянной. Однако из-за сеточной погрешности она представляет собой некоторой профиль, зависящий от z. Вычислим норму L_2 этого профиля. Полученное число есть расчетный коэффициент отражения при данном λ.

Аналогично определялся расчетный коэффициент прохождения. Прошедшая волна E_{transm} есть численное решение исходной задачи правее крайней правой пластины.

Сеточная погрешность расчета контролировалась по методу сгущения сеток и составляла не более $\sim 0.1\%$.

2° В спектрах хорошо видна запрещенная зона в диапазоне длин волн $850 \div 1050$ нм. В ней отражение является почти полным: ФК работает как зеркало. Слева и справа от запрещенной зоны видны максимумы прохождения $T \approx 1$; в них отражение практически отсутствует $R \approx 0$.

Отметим, что в эксперименте измерение отражения при нормальном падении достаточно проблематично. Расчет отражения на рис. 6.15 выполнен только для верификации бикомпактной схемы.

3° Для сравнения такой же расчет был выполнен с помощью метода матриц рассеяния [313]. Использована реализация этого метода, опубликованная в [314]. Все среды являются изотропными, поэтому метод [313] дает те же результаты, что и метод Берремана [315]. Напомним, что для стационарной задачи методы рассеяния дают точные значения коэффициентов отражения и прохождения; т.е. соответствующие спектры являются точными.

Спектры, найденные матричным методом, также приведены на рис. 6.15, 6.16. Видно, что с графической точностью спектры, найденные матричным ме-

тодом и бикомпактной схемой, совпадают. Положения экстремумов отличаются менее, чем на 0.1%, что соответствует точности расчета по бикомпактной схеме. Это убедительно подтверждает правильность бикомпактной схемы.

Экспериментальный спектр. Для этого ФК доступен экспериментально измеренный спектр прохождения при нормальном падении. Эти измерения выполнены производителем данного ФК [316] и любезно предоставлены автору Бессоновым и Попковой, входящими в состав коллектива под руководством Федянина на кафедре квантовой электроники физического факультета МГУ им. М.В. Ломоносова. Этот спектр показан на рис. 6.17.

Отметим, что этот спектр отличается от расчетного спектра, приведенного на рис. 6.16. Максимумы прохождения оказываются сглаженными и не достигают ста процентов.

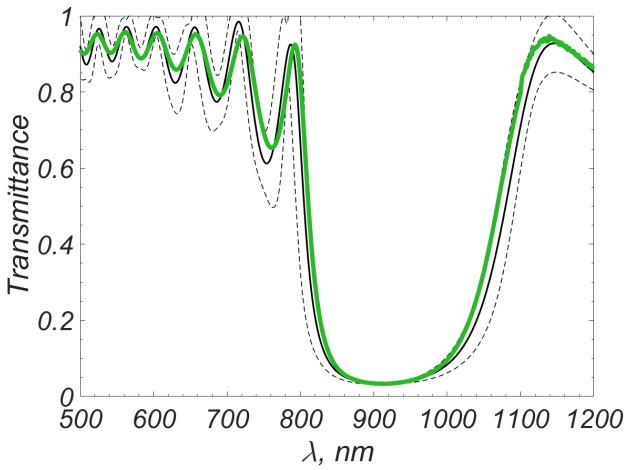

Рис. 6.17. Спектр прохождения ФК на рис. 6.12. Жирная линия – эксперимент [316]. Тонкая линия – бикомпактная схема и виртуальный эксперимент. Штриховая линия – границы доверительного интервала по методу виртуального эксперимента (соответствуют двум стандартным уклонениям).

Виртуальный эксперимент. 1° Изготовление оптических покрытий – сложный технологический процесс. Из-за этого геометрические параметры рассеивателя (например, толщины слоев ФК) могут флуктуировать. Эти флуктуации составляют несколько процентов от размера рассеивателя. Они вносят существенные искажения в электромагнитный отклик конструкции. На-

пример, колебания толщины слоя ФК влияют на интерференцию отраженных волн внутри ФК. Это приводит к размытию резких минимумов и максимумов в спектрах отражения и пропускания. Физическая природа этого фактора хорошо известна [317].

Численное моделирование этого фактора на примере дифракционных решеток проводилось коллективом Сойфера [289]. Были выполнены расчеты дифракции на одиночном деформированном элементе решетки, и было исследовано влияние величины деформации на спектр отражения. Однако предположение о том, что все элементы рассеивателя деформированы одинаково, является слишком грубым. В реальном оптическом покрытии геометрические размеры элементов флуктуируют относительно средних значений, причем детальная реализация этих флуктуаций зачастую неизвестна. Это затрудняет применение методов типа преобразования координат и разложения преобразованного поля. Требуется универсальная процедура расчета, которая позволяет учитывать описанный фактор.

2° В данной работе предлагается ***метод виртуальноого эксперимента***. Пусть один из параметров задачи A известен с ошибкой σ_A. Будем считать этот параметр нормально распределенной случайной величиной с математическим ожиданием A и стандартным отклонением σ_A. Электромагнитный отклик структуры является функцией этой случайной величины. Поэтому результаты расчета необходимо усреднить по распределению $P(A)$ значения A. Например, средний коэффициент прохождения равен

$$\langle T\rangle = \int_{-\infty}^{\infty} T(a)P(a)da, \quad P(a) = \frac{1}{\sigma_A\sqrt{2\pi}}\exp\left\{-\frac{(a-A)^2}{2\sigma_A^2}\right\}. \tag{6.23}$$

Одновременно можно найти доверительный интервал

$$\sigma_T = \sqrt{\langle T^2\rangle - \langle T\rangle^2}. \tag{6.24}$$

Величина (6.24) есть оценка погрешности расчета, обусловленная погрешностью задания параметра A.

Параметр A может иметь различную природу: геометрические параметры элементов структуры, неоднородности плотности, приводящие к локальным флуктуациям показателя преломления, угловое расхождение падающего пучка и т.д. Метод виртуального эксперимента позволяет учитывать эти типы возмущений единообразно. Мы рекомендуем использовать его во всех прикладных расчетах как по численным, так и по аналитическим методам.

В случае неидеального фотонного кристалла A есть вектор, содержащий толщины всех слоев. Описанную процедуру можно интерпретировать следующим образом. Вся поверхность ФК разбивается на участки, соответствующие фиксированным толщинам слоев. В рамках одного участка толщина каждого слоя постоянна и изменяется при переходе к соседним участкам. Сумма откликов от таких участков определяет отклик ФК в целом. Это напоминает определение интеграла Лебега по схеме Даниэля.

Распределение $P(A)$ является многомерным; оно равно произведению соответствующего числа одномерных гауссовых распределений. Поэтому квадратуры (6.23), (6.24) имеют большую размерность. Их удобно вычислять методом Монте-Карло. В задачах нанофотоники этот подход ранее не предлагался.

Результаты моделирования. 1º На рис. 6.17 показан спектр, рассчитанный по бикомпактной схеме с методом виртуального эксперимента. В расчетах учитывались колебания толщины с типичным значением ± 4.5 нм для всех слоев ФК. Это согласуется с известными оценками, полученными с помощью эллипсометрии. Для расчетной кривой представлен доверительный интервал, то есть оценка погрешности из-за неточности задания толщин слоев. Эта оценка была получена с помощью виртуального эксперимента и соответствует двум стандартным отклонениям.

2º Видно, что спектр, рассчитанный по бикомпактной схеме и методу виртуального эксперимента, согласуется с экспериментальным спектром в пределах 2-5% в широком диапазоне длин волн 500-1200 нм. Это соответствует точности самого эксперимента.

Также видно, что ширина доверительного интервала по методу виртуального эксперимента сопоставима с расхождением расчетного и экспериментального спектров. Поэтому флуктуации толщин слоев ФК являются важным фактором, и их необходимо учитывать в расчетах.

3° Предложенные методы обеспечивают высокую количественную точность расчета реальных нестационарных задач фотоники. Это особенно важно при численных исследованиях сложных эффектов, таких как долгоживущие связанные состояния (поверхностные волны различных типов), а также при разработке новых устройств интегральной фотоники.

6.2. Нестационарные задачи

В нестационарном случае были проведены расчеты трех тестовых задач: распространение излучения в однородной среде (в том числе диспергирующей), распространение излучения через границу раздела, излучение поверхностных токов). Временная развертка падающего импульса может быть задана произвольно. Для определенности в задачах этого пункта он считается гауссовым

$$\mathcal{E}(t) = \exp\left(-\frac{t^2}{2W^2} - i\omega^0 t\right), \quad \omega^0 = \frac{3\pi}{2}\pi, \quad W = \frac{\pi}{4}. \tag{6.25}$$

Напомним, что все величины считаются обезразмеренными. Ширина импульса составляет $\approx \pi$ на половине высоты и $\approx 3\pi$ по основанию. Временная развертка импульса представлена на рис. 6.18, а его спектр – на рис. 6.19.

6.2.1. Однородная среда

1° Рассмотрим первую задачу из раздела 6.1 в нестационарной постановке. Пусть в однородной среде с материальными параметрами $\varepsilon = 1 + 5i$, $\mu = 1$ распространяется импульс (6.25).

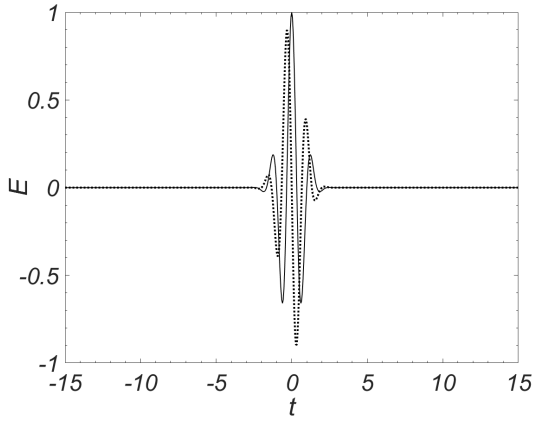

Рис. 6.18. Временная развертка падающего импульса. Сплошная линия – $\mathrm{Re}\, E$, пунктир – $\mathrm{Im}\, E$

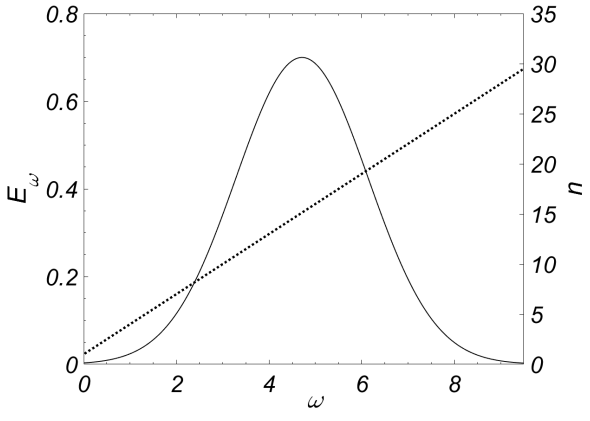

Рис. 6.19. Спектр падающего импульса (сплошная линия) и зависимость показателя преломления от частоты в задаче о диспергирующей среде (пунктир).

2° Построим точное решение этой задачи. Разложим падающий импульс $\mathcal{E}(t)$ в интеграл Фурье

$$E_\omega = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathcal{E}(t) e^{i\omega t} dt \tag{6.26}$$

Решим задачу (5.21) – (5.24) для (5.23) $E^0 = E_\omega$, $E^a = 0$. Это решение имеет вид (6.5). Выполним обратное преобразование Фурье

$$\begin{gathered} E_x(z,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} E_\omega e^{i\tilde{k}z} e^{-i\omega t} d\omega = \mathcal{E}\left(t - \sqrt{\varepsilon\mu}\frac{z}{c}\right), \quad \tilde{k} = \frac{\omega\sqrt{\varepsilon\mu}}{c}, \\ H_y(z,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sqrt{\frac{\varepsilon}{\mu}} E_\omega e^{i\tilde{k}z} e^{-i\omega t} d\omega = \sqrt{\frac{\varepsilon}{\mu}} \mathcal{E}\left(t - \sqrt{\varepsilon\mu}\frac{z}{c}\right). \end{gathered} \tag{6.27}$$

Это решение известно из учебников.

3° Расчет проводился по нестационарной бикомпактной схеме. Полученные погрешности приведены на рис. 6.20. Видно, что, во-первых, погрешности, найденные сравнением с точным решением, совпадают с оценками по Ричардсону, и во-вторых, фактически порядок точности равен 2 и совпадает с теоретическим.

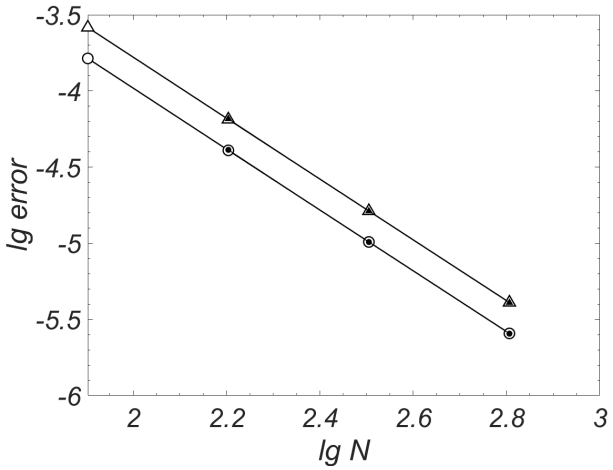

Рис. 6.20. Погрешность решения в нестационарной задаче о среде с поглощением. Обозначения соответствуют рис. 6.2

6.2.2. Диспергирующая среда

1° Рассмотрим предыдущую задачу в усложненной постановке. Пусть теперь среда является диспергирующей, то есть величины $\varepsilon(\omega)$ и $\mu(\omega)$ зависят от частоты.

2° Точное решение по-прежнему выражается квадратурой (6.27), однако класс $\varepsilon(\omega)$ и $\mu(\omega)$, при которых ее удается вычислить ее в элементарных функциях, весьма узок.

Положим $\varepsilon(\omega) = \mu(\omega) = A_1\omega + A_0$. Тогда обе квадратуры в (6.27) берутся через интеграл Пуассона

$$E_x(z,t) = \frac{W}{\sqrt{2a}}\exp\left(\frac{b^2}{4a} - c\right), \quad H_y(z,t) = E_x(z,t)$$
$$a = \frac{W^2}{2} - i\frac{A_1}{c}z, \quad b = \omega^0 W^2 + i\left(\frac{A_0 z}{c} - t\right), \quad c = \frac{1}{2}(\omega^0 W)^2. \tag{6.28}$$

Этот случай представляет большую ценность, поскольку обе компоненты решения выражаются явно. По-видимому, это решение является новым.

В конкретном расчете были выбраны значения $A_1 = 3$, $A_0 = 1$. Соответствующая зависимость $n(\omega) = \sqrt{\varepsilon(\omega)\mu(\omega)}$ показателя преломления от частоты приведена на рис. 6.19 (правая ось ординат). Видно, что на отрезке частот, со-

держащихся в импульсе (6.25), показатель преломления изменяется почти в 30 раз: от $n = 1$ до $n \approx 30$. Поэтому данный тест достаточно представителен.

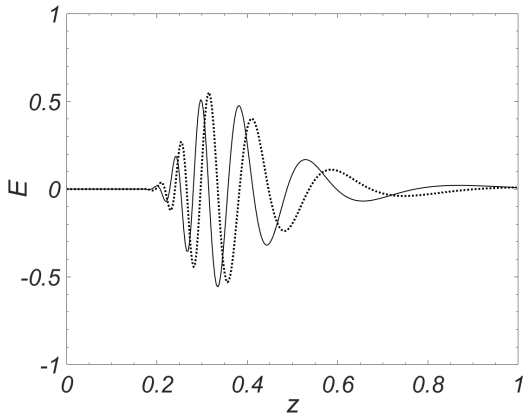

Рис. 6.21. Пространственная развертка решения в задаче о диспергирующей среде. Обозначения соответствуют рис. 6.18

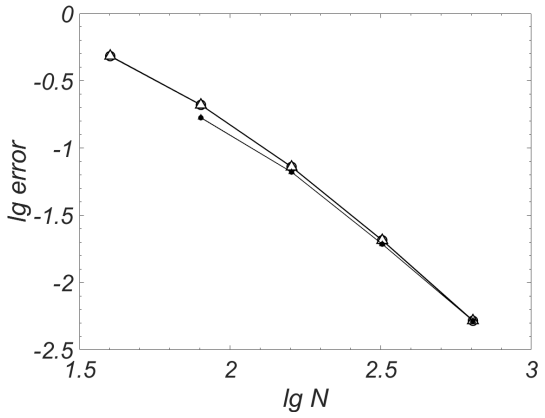

Рис. 6.22. Погрешность решения в задаче о диспергирующей среде. Обозначения соответствуют рис. 6.2

3° Проанализируем характер решения. На рис. 6.21 показана зависимость E_x и H_y от z при фиксированном t. Видно, что высокочастотные компоненты, для которых $n(\omega)$ велико, распространяются медленно, а низкочастотные их заметно обгоняют. В результате импульс расплывается, причем несимметрично.

За время T область локализации импульса увеличивается на $cT(1/\min n - 1/\max n)$, это наибольший из пространственных масштабов задачи. Наименьший масштаб есть длина волны самой высокочастотной компоненты $\min\lambda = 2\pi c/\max(\omega n)$. Поэтому отношение наибольшего пространственного масштаба к наименьшему в данной задаче равно

$$\frac{cT(1/\min n - 1/\max n)}{2\pi c/\max(\omega n)} \approx 46T. \tag{6.29}$$

Видно, что увеличивая T, разномасштабность (6.29) можно сделать сколь угодно большой. Поэтому задачу следует относить к жестким. Это является дополнительным преимуществом предложенного точного решения, поскольку оно оказывается хорошим тестом для алгоритмов адаптивного выбора шага по пространству и времени. Особенность этого теста заключается в том, что область,

где нужно сгущать пространственную сетку, перемещается с течением времени, и ее положение заранее неизвестно.

В проведенных расчетах полное время равнялось $T = 30$, и величина (6.29) достигала ~ 1400. Это достаточно высокая жесткость. Расчет проводился с постоянным шагом по пространству и времени. Поэтому требовались достаточно подробные сетки.

4° На рис. 6.22 показаны полученные погрешности. Видно, что при сгущении сеток погрешности убывают в соответствии со вторым порядком точности, и оценки по методу Ричардсона хорошо согласуются с погрешностями, определенными непосредственно сравнением численного решения с точным. Этот тест убедительно подтверждает правильность предложенного метода учета дисперсии.

5° Отметим еще два случая, когда хотя бы часть решения удается выразить явно. Во-первых, пусть $\mu = 1$, и $\varepsilon(\omega)$ представимо в виде отношения квадратов некоторых полиномов

$$\begin{gathered} \varepsilon(\omega) = \left(\frac{P_1(\omega)}{P_2(\omega)} \right)^2, \\ P_1(\omega) = A_n \omega^n + A_{n-1}\omega^{n-1} + \ldots + A_0, \\ P_2(\omega) = B_m \omega^m + B_{m-1}\omega^{m-1} + \ldots + B_0. \end{gathered} \tag{6.30}$$

Тогда интеграл для $H_y(z,t)$ берется через вычеты в нулях полинома P_2. Однако вычислить квадратуру для $E_x(z,t)$ не удается. Такой тест менее информативен, чем описанный выше. Тем не менее, диэлектрическую проницаемость многих реальных материалов можно описать равенством (6.30). Поэтому данный случай также представляет интерес.

Во-вторых, гауссова огибающая является моделью. Вместо нее можно использовать огибающую следующего вида:

$$\mathcal{E}(t) = \frac{A}{B + t^{2k}} \tag{6.31}$$

Здесь k – натуральное число. Качественное поведение (6.31) аналогично (6.25). В этом случае квадратуры (6.27) можно вычислить через вычеты при произвольном законе дисперсии $\varepsilon(\omega)$, $\mu(\omega)$.

Насколько нам известно, в литературе эти случаи не описаны.

6.2.3. Граница раздела диэлектриков

1° Рассмотрим эту задачу в нестационарной постановке: пусть слева на границу раздела падает импульс вида (6.25).

2° Разложим падающий импульс в интеграл Фурье и для каждой спектральной амплитуды E_ω решим стационарную задачу (5.21) – (5.24). Точное решение последней имеет вид (6.10) – (6.13), где $E^0 = E_\omega$. Обратное преобразование Фурье дает точное решение нестационарной задачи: при $z < b$ решение состоит из падающей и отраженной волн

$$\begin{aligned} E_x &= \mathcal{E}\left(\sqrt{\varepsilon_1\mu_1}\frac{z}{c} - t\right) + C_2\mathcal{E}\left(\frac{2b-z}{c}\sqrt{\varepsilon_1\mu_1} - t\right), \\ H_y &= \sqrt{\frac{\varepsilon_1}{\mu_1}}\mathcal{E}\left(\sqrt{\varepsilon_1\mu_1}\frac{z}{c} - t\right) - \sqrt{\frac{\varepsilon_1}{\mu_1}}C_2\mathcal{E}\left(\frac{2b-z}{c}\sqrt{\varepsilon_1\mu_1} - t\right); \end{aligned} \tag{6.32}$$

а при $z > b$ – из прошедшей

$$\begin{aligned} E_x &= C_3\mathcal{E}\left(\frac{b}{c} + \sqrt{\varepsilon_2\mu_2}\frac{z-b}{c} - t\right), \\ H_y &= C_3\sqrt{\frac{\varepsilon_2}{\mu_2}}\mathcal{E}\left(\frac{b}{c} + \sqrt{\varepsilon_2\mu_2}\frac{z-b}{c} - t\right). \end{aligned} \tag{6.33}$$

Найти это решение в литературе не удалось. По-видимому, оно является новым.

3° Выберем такие же материальные параметры, как в п. 6.1: $\varepsilon_1 = \mu_1$, $\varepsilon_2 = 10$, $\mu_2 = 2$. Расчет проводился по нестационарной схеме, соответствующие погрешности приведены на рис. 6.23. Видно, что, во-первых, погрешности, найденные сравнением с точным решением, совпадают с оценками по Ричардсону, и во-вторых, фактически порядок точности равен 2 и совпадает с теоретическим.

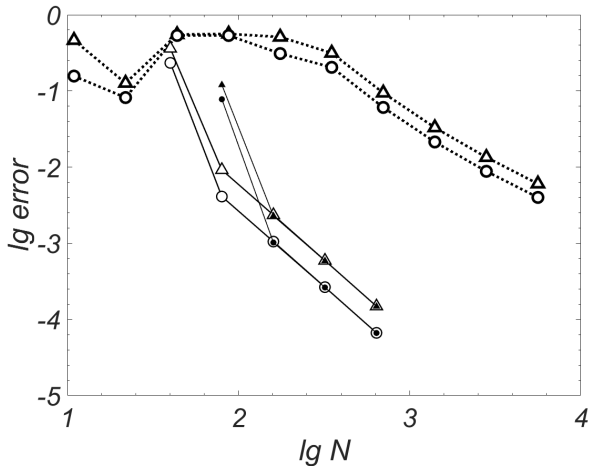

Рис. 6.23. Погрешность решения в нестационарной задаче о границе раздела диэлектриков. Сплошные линии – расчет по бикомпактной схеме, пунктир – по методу FDTD, другие обозначения соответствуют рис. 6.2

4° Для сравнения проводился расчет нестационарной задачи по схеме «с перешагиванием» метода FDTD. Граница раздела расположена в целом узле (напомним, что в этом методе к целым узлам относится поле E_x, а к полуцелым – поле H_y). Соответствующие погрешности приведены на рис. 6.23. Видно, что они существенно хуже, чем у бикомпактной схемы, а порядок точности только 1-й (вместо теоретического 2-го).

Причина снижения порядка точности у метода FDTD заключается в том, что схема «с перешагиванием» некомпактна: шаблон занимает 1.5 шага по пространству и по времени. Поэтому, как бы ни была расположена граница раздела (в целом узле или в полуцелом), обеспечить достаточную гладкость решения внутри шаблона не удается. Это приводит к большой немонотонности решения (то есть появлению сильных нефизичных осцилляций). Величина этих осцилляций составляет $O(h+\tau)$. Это и приводит к падению точности вблизи границ раздела.

5° В литературе предлагался ряд подходов, основанных на «размывании» границы раздела, то есть замене скачка показателя преломления некоторым сглаженным профилем. Известно, что это вносит физическую погрешность в расчетное отражение. Приведем пример такого расчета.

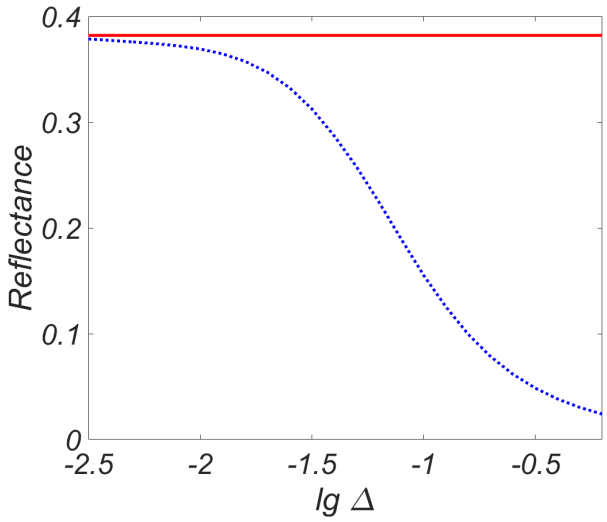

Рис. 6.24. Отражение от сглаженной границы раздела. Обозначения – см. текст.

Был выполнен расчет тестовой задачи со сглаженной границей раздела. Пусть исходная (скачкообразная) граница раздела расположена между средами с $\varepsilon_l = 1$ (слева) и $\varepsilon_r = 5$ (справа). Зададим сглаживание $\varepsilon(x)$

$$\varepsilon(x) = \varepsilon_l + (\varepsilon_r - \varepsilon_l)\mathrm{th}[(x - x_0)/\Delta], \quad x \in [0, 1]. \tag{6.34}$$

Здесь Δ - характерная ширина размытия, $x_0 = 0.5$ – положение границы раздела.

Падающая волна движется слева направо. Коэффициент отражения R вычислим по процедуре, описанной в п. 6.1.5. На рис. 6.24 показана зависимость R от Δ (пунктир). Масштаб оси абсцисс логарифмический, чтобы охватить широкий диапазон Δ. Для сравнения приведено отражение от исходной (т.е. скачкообразной) границы раздела (пунктир), которое известно из теории $R_{\mathrm{th}} = (\sqrt{\varepsilon_r} - \sqrt{\varepsilon_l})/(\sqrt{\varepsilon_r} + \sqrt{\varepsilon_l}) = 0.382\ldots$

Видно, что сглаживание границы раздела уменьшает отражение; и чем шире размытие, тем сильнее оно влияет на отражение. Узкое размытие $\Delta \sim 0.01$ лишь слегка искажает отражение. Однако среда все еще сильно неоднородна и представляет серьезную проблему для схемы Йе, поскольку амплитуда нефизичных осцилляций все еще велика [301]. Достаточно широкое размытие $\Delta \sim 0.1$

ослабляет немонотонность, но вносит катастрофически большие искажения в отражение.

6.2.4. Граница раздела с токами

1° В нестационарном случае зададим поверхностный ток на границе раздела в виде (6.25).

2° Построим точное решение этой задачи. Разложим поверхностный ток в интеграл Фурье и обозначим Фурье-образ j_1 через j_ω. Решим стационарную задачу (5.21) – (5.24) при $E^0 = E^a = 0$ и $j_1 = j_\omega$. Ее точное решение имеет вид (6.10), (6.11), (6.17), (6.18).

Выполним обратное преобразование Фурье и получим точное решение нестационарной задачи. При $z < b$ оно имеет вид

$$\begin{gathered} E_x(z,t) = \frac{4\pi}{c}\frac{1}{\sqrt{\varepsilon_1/\mu_1}+\sqrt{\varepsilon_2/\mu_2}} j_1\left(\frac{b-z}{c}\sqrt{\varepsilon_1\mu_1} - t\right), \\ H_y(z,t) = -\sqrt{\frac{\varepsilon_2}{\mu_2}} E_x(z,t). \end{gathered} \tag{6.35}$$

При $z > b$ нужно поменять местами индексы 1 и 2 и изменить знак разности $(b-z) \to (z-b)$

$$\begin{gathered} E_x(z,t) = \frac{4\pi}{c}\frac{1}{\sqrt{\varepsilon_1/\mu_1}+\sqrt{\varepsilon_2/\mu_2}} j_1([z-b]\sqrt{\varepsilon_2\mu_2}/c - t), \\ H_y(z,t) = -\sqrt{\frac{\varepsilon_2}{\mu_2}} E_x(z,t). \end{gathered} \tag{6.36}$$

Решение (6.35), (6.36) является новым.

3° Погрешности расчета по нестационарной бикомпактной схеме приведены на рис. 6.25. Видно, что, начиная со второй сетки ($N \approx 80$), погрешность убывает в соответствии со вторым порядком точности. Величины погрешностей, вычисленные по методу Ричардсона, практически совпадают с погрешностями, вычисленными по разности численного и точного решений.

4° Для сравнения был проведен расчет нестационарной задачи по явной схеме «с перешагиванием» метода FDTD. При этом расчет велся по однородной

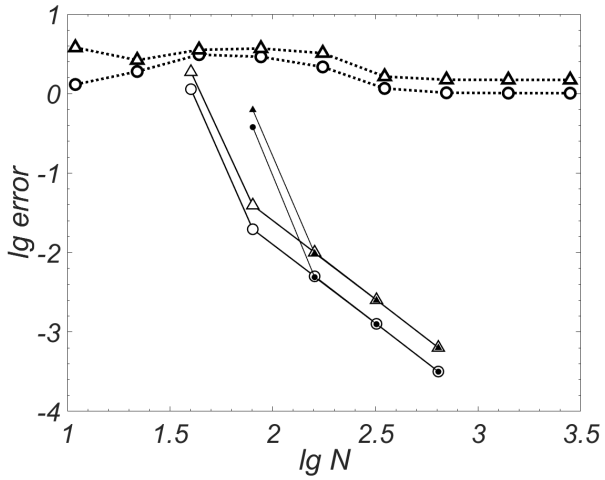

Рис. 6.25. Погрешность решения в нестационарной задаче о границе раздела с поверхностными токами. Обозначения соответствуют рис. 6.23

схеме, то есть без явного выделения особенности. Соответствующие погрешности даны на рис. 6.25. Видно, что этот метод вообще не может обеспечить сходимость: при уменьшении шага погрешность не убывает.

5° Таким образом, приведенные примеры расчетов убедительно показывают преимущества предложенного метода по сравнению с известными подходами.

6.3. Основные результаты главы

1. Построены физически содержательные и сложные для расчета тестовые примеры. С их помощью показано, что разностные методы, предложенные в главе 5, кардинально превосходит существующие подходы по точности и надежности.
2. Проведены расчеты задачи о нормальном падении плоской волны на одномерный фотонный кристалл. Для учета флуктуаций толщин слоев образца предложен новый подход, названный методом виртуального эксперимента. Рассчитан спектр прохождения через фотонный кристалл. Результаты расчета хорошо согласуются с ранее измеренным спектром в пределах экспериментальных погрешностей 2-5%.

7. Метод оптических путей

7.1. Наклонное падение плоской волны на плоский рассеиватель

7.1.1. Рассеяние монохроматического излучения плазмонными структурами

1° Рассмотрим многослойную плоско-параллельную структуру из п. 5.1.1. Пусть часть пластин является диэлектрическими, часть – проводящими (т.е. проводниками или полупроводниками).

2° Пусть на эту структуру наклонно падает плоская линейно поляризованная электромагнитная волна частоты ω. Ось z направим перпендикулярно пластинам. Ось x направим так, чтобы волновой вектор лежал в плоскости Oxz. Угол падения (то есть угол между волновым вектором падающей волны и нормалью к пластинам) обозначим через α.

Как известно, при наклонном падении возможны две поляризации волны. Волна называется s-поляризованной, если вектор $\mathbf{E}$ перпендикулярен плоскости, образованной волновыми векторами падающей и отраженной волн. Тогда вектор $\mathbf{E}$ имеет только y-компоненту $\mathbf{E} = \{0, E_y, 0\}$, а вектор $\mathbf{H}$ – x- и z-компоненты $\mathbf{H} = \{H_x, 0, H_z\}$. Такая поляризация проиллюстрирована на рис. 7.1, где для простоты приведена одна граница раздела. Если вектор $\mathbf{E}$ лежит в плоскости, образованной волновыми векторами падающей и отраженной волн, то волна называется p-поляризованной (см. рис. 7.2). Тогда вектор $\mathbf{E}$ имеет только x- и z-компоненты $\mathbf{E} = \{E_x, 0, E_z\}$, а вектор $\mathbf{H}$ – только y-компоненту $\mathbf{H} = \{0, H_y, 0\}$.

При этом если падающая волна имеет s-поляризацию, то отраженная и прошедшая волны также будут s-поляризованными; аналогично в случае p-поляризации. Мы будем считать, что падающая волна является s- либо p-поляризованной. Именно этот случай представляет интерес для приложений.

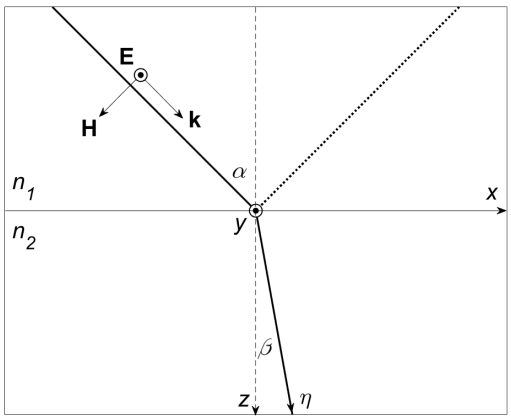

Рис. 7.1. S-поляризованная волна. Жирная линия – направление распространения прямой волны (лучевая траектория). Пунктир – направление распространения отраженной волны. Тонкая линия – граница раздела сред. Штриховая линия – нормаль к границе раздела.

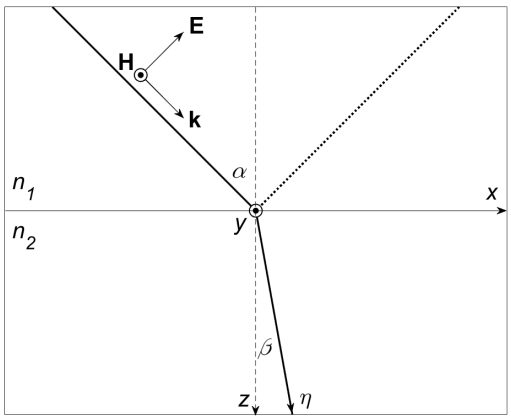

Рис. 7.2. P-поляризованная волна. Обозначения соответствуют рис. 7.1.

3° Пусть внешние объемные токи отсутствуют $\mathbf{J}^{\text{ext}} = 0$. Падающее излучение индуцирует объемные токи $\mathbf{J}_q^{\text{ind}} = \sigma_q \mathbf{E}_q$. Они направлены так же, как вектор $\mathbf{E}$. Эти токи излучают волны, которые интерферируют с падающей, отраженной и прошедшей волнами. При этом могут формироваться различные связанные состояния (см. п. 5.1.3).

4° Будем считать, что материал пластин может быть неоднородным из-за нагрева токами и падающим излучением. При этом будем считать, что ε_q, μ_q, σ_q зависят только от z и не зависят от x, y.

Данная задача возникает в плазмонике [276–278].

5° Математическая постановка этой задачи включает уравнения Максвелла (5.4), условия сопряжения (5.5) при $j^{\text{surf}} = 0$ и двумерный аналог условий излучения на границах расчетной области.

Сформулируем условия излучения для данной задачи. Выделим участок рассеивателя, ограниченный плоскостями $x = 0$ и $x = d > 0$. Для компактности записи введем следующие обозначения:

$$A_q^{\pm} = \frac{\partial}{\partial x} \pm i(\tilde{k}_q)_x, \quad B_q^{\pm} = \frac{\partial}{\partial z} \pm i(\tilde{k}_q)_z. \tag{7.1}$$

Здесь $(\tilde{k}_q)_x = \tilde{k}_q \sin\alpha_q$ и $(\tilde{k}_q)_z = \tilde{k}_q \cos\alpha_q$ – проекции волнового вектора в q-й пластине на оси x и z соответственно. Значения $1 \leqslant q \leqslant Q-1$ соответствуют пластинам рассеивателя, $q = 0$ – среде при $z < 0$, $q = Q$ – среде при $z > a$. Напомним, что волновое число в q-й пластинке равно $\tilde{k}_q = \omega\sqrt{\varepsilon_q \mu_q}/c$.

Оператор $A_q^+ B_q^+$ позволяет выделить падающую волну $\sim \exp\{i(\tilde{k}_q)_x x + i(\tilde{k}_q)_z z\}$ в q-й среде и задать ее амплитуду, не затрагивая амплитуду отраженной волны $\sim \exp\{i(\tilde{k}_q)_x x - i(\tilde{k}_q)_z z\}$.

Таким образом, условия излучения принимают следующий вид:

$$A_q^+ B_q^+ \mathbf{E} = -4(\tilde{k}_q)_x (\tilde{k}_q)_z \mathbf{E}_q^0 e^{i(\tilde{k}_q)_z z}, \quad x = 0, \quad \xi_{q-1} < z < \xi_q, \tag{7.2}$$

$$A_0^+ B_0^+ \mathbf{E} = -4(\tilde{k}_0)_x (\tilde{k}_0)_z \mathbf{E}_0^0 e^{i(\tilde{k}_0)_z z}, \quad z = 0, \quad 0 < x < d, \tag{7.3}$$

$$A_q^- B_q^+ \mathbf{E} = 0, \quad x = d, \quad \xi_{q-1} < z < \xi_q, \tag{7.4}$$

$$A_Q^+ B_Q^- \mathbf{E} = 0, \quad z = a, \quad 0 < x < d. \tag{7.5}$$

Здесь $\mathbf{E}_q^0$ – амплитуда волны, падающей на границу раздела $z = \xi_q$ из среды q в среду $q+1$. Амплитуда $\mathbf{E}_{q+1}^0$ связана с $\mathbf{E}_q^0$ коэффициентами Френеля. Амплитуда $\mathbf{E}_0^0$ соответствует излучению, падающему на границу рассеивателя $z = 0$. Условия (7.4) показывают, что падающая и отраженная волны свободно проходят через границу $x = d$, а условие (7.5) – что прошедшая волна свободно проходит через границу $z = a$.

7.1.2. Рассеяние монохроматического излучения оптическими структурами

1° Пусть в задаче п. 7.1.1 все пластины выполнены из диэлектрических материалов и прозрачны для падающего излучения (то есть величина $\operatorname{Im}\varepsilon \ll 1$

мала). Пусть все объемные и поверхностные токи отсутствуют $J_q^{\text{ext}} = J_q^{\text{ind}} = 0$, $j_q^{\text{ext}} = j_q^{\text{ind}} = 0$.

2° Пусть на эту структуру наклонно падает плоская линейно поляризованная электромагнитная волна частоты ω. Направление осей координат выберем так же, как в задаче п. 7.1.1. Возможны две поляризации падающей волны, они показаны на рис. 7.1, 7.2. Падающее излучение частично отражается от структуры и частично проходит через нее. Внутри структуры могут образовываться связанные состояния, например, поверхностные волны Блоха [318, 319].

3° Будем считать пластинки пространственно однородными, т.е. внутри каждой пластины материальные параметры ε_q, μ_q постоянны. При этом предполагается, что нагрев пластин падающим излучением пренебрежимо мал.

4° В литературе данную задачу принято называть оптической. Она часто встречается во многих технических приложениях (см. п. 5.1.4).

5° Математическая постановка задачи включает уравнения Максвелла (5.4), условия сопряжения (5.5) и двумерные условия излучения (7.2) – (7.5).

7.1.3. Рассеяние электромагнитного импульса плазмонными структурами

1° Пусть в задаче п. 7.1.1 на рассеиватель падают не монохроматическая волна, а волновой пакет

$$\mathbf{f}^0(\zeta) = \mathbf{E}^0(\zeta) \exp\left(-i\omega^0\zeta\right), \quad \zeta = t - \frac{(\tilde{\mathbf{k}}^0\mathbf{r})}{\tilde{k}^0 c} \tag{7.6}$$

с несущей частотой ω^0 и заданной огибающей $\mathbf{E}^0$. Здесь $\tilde{\mathbf{k}}^0/\tilde{k}^0$ – единичный вектор в направлении распространения волны, $k^0 = \omega^0/c$. Пусть импульс (7.6) является плоским и линейно поляризованным, а огибающая – финитной. Оси координат выберем так же, как в задаче (7.1.1).

Падающее излучение индуцирует объемные токи $\mathbf{J}_q^{\text{ind}} = \sigma_q \mathbf{E}_q$, которые переизлучают линейно поляризованные импульсы.

2° Пусть показатель преломления и проводимость зависят от координаты z, но медленно меняются со временем. Среды имеют частотную дисперсию, но пространственная дисперсия пренебрежимо мала. Условия, когда эти предположения верны, сформулированы в п. 5.1.5.

3° Описанная задача возникает при рассмотрении сверхбыстрых процессов в плазмонных структурах (см. п. 5.1.6); она имеет большое значение для практических приложений. Данная задача является обобщением задачи п. 7.1.1.

4° Приведем математическую постановку задачи. Она содержит уравнения Максвелла (5.14), начальные условия (5.15), условия сопряжения (5.5) при $j^{\text{surf}} = 0$ и нестационарные двумерные условия излучения на границах $x = 0$, $x = d$, $z = 0$, $z = a$.

Сформулируем нестационарный аналог условий (7.2) – (7.5). Для этого в них нужно заменить $i\omega \to \partial/\partial t$ и подставить в правую часть импульс (7.6). Обозначим

$$C_q^{\pm} = \frac{\partial}{\partial x} \mp \frac{\sin\alpha_q}{c}\frac{\partial}{\partial t}, \quad D_q^{\pm} = \frac{\partial}{\partial z} \mp \frac{\cos\alpha_q}{c}\frac{\partial}{\partial t}. \tag{7.7}$$

Тогда нестационарные условия излучения, описывающие падение волны на рассеиватель, принимают вид

$$C_q^+ D_q^+ \mathbf{E} = -4\frac{(\omega^0)^2}{c^2}\sin\alpha_q \cos\alpha_z \frac{d^2\mathbf{f}^q}{d\zeta^2}, \quad \zeta = t - \frac{(\tilde{\mathbf{k}}^0\mathbf{r})}{\tilde{k}^0 c} \tag{7.8}$$

при $x = 0$, $\xi_{q-1} < z < \xi_q$ и при $z = 0$, $0 < x < d$. Здесь $\mathbf{f}^q$ – профиль падающего импульса в q-й пластине. Условия излучения, описывающие уход на бесконечность прошедшей волны, записываются следующим образом:

$$C_q^- D_q^+ \mathbf{E} = 0, \quad x = d, \quad \xi_{q-1} < z < \xi_q, \tag{7.9}$$

$$C_Q^+ D_Q^- \mathbf{E} = 0, \quad z = a, \quad 0 < x < d. \tag{7.10}$$

Условия (7.8) – (7.10) интерпретируются аналогично (7.2) – (7.5)

7.1.4. Рассеяние электромагнитного импульса оптическими структурами

1° Пусть в задаче п. 7.1.2 на рассеиватель падает волновой пакет (7.6). Оси координат выберем так же, как в п. 7.1.2. Будем считать, что материальные параметры ε_q, μ_q постоянны в пределах каждой пластины. Имеет место частотная дисперсия, причем пространственная дисперсия считается пренебрежимо малой.

2° Данная задача имеет большое значение для исследования динамики связанных состояний в полностью диэлектрических структурах (см. п. 5.1.7).

3° Математическая постановка задачи включает уравнения Максвелла (5.17), начальные условия (5.18), условия сопряжения (5.19) и условия излучения (7.8) – (7.10).

7.2. Задача в интегральной форме

Стационарная задача включает интегральные уравнения Максвелла (5.21), (5
условия сопряжения (5.24) при $j^{\text{surf}} = 0$ и условия излучения (7.2) – (7.5).

Нестационарная задача состоит из уравнений Максвелла (5.25), (5.26), условий сопряжения (5.28) при $j^{\text{surf}} = 0$, начальных условий (5.29) и условий излучения (7.8) – (7.10).

Очевидно, эти постановки содержат задачи п. 7.1.1 – 7.1.4 как частные случаи.

7.3. Известные методы

Задачи, описанные в п. 7.1.1 – 7.1.4, являются двумерными. Обзор методов для них дан в приложении 2.

Для задачи п. 7.1.2 наиболее эффективны методы типа матриц рассеяния. Для задачи п. 7.1.1 при отсутствии поверхностных токов наиболее работоспо-

собны, по-видимому, методы конечных и граничных элементов. Для задач, в которых есть поверхностные токи, методы не разработаны.

Для нестационарных задач п. 7.1.3, 7.1.4 при отсутствии поверхностных токов применимы только методы конечных разностей и конечных элементов во временной области. Как отмечалось ранее, эти методы обладают рядом принципиальных недостатков. Методы, позволяющие рассчитывать решения с сильным разрывом, не предложены.

7.4. Оптические пути

7.4.1. Снижение размерности многомерных задач

Задачи для уравнений математической физики со многими переменными представляют большую вычислительную сложность. Существует, однако, ряд приемов, которые позволяют снизить размерность задачи. Это аналитико-численные алгоритмы, как правило, основанные на физических упрощениях задачи. Одним из них является автомодельная замена переменных, что превращает уравнение в частных производных в ОДУ. Построение таких замен можно считать скорее искусством; общие алгоритмы отсутствуют.

Другим хорошо известным способом является метод характеристик для гиперболических задач. Интегрирование исходного уравнения вдоль характеристики можно трактовать как задачу сниженной размерности.

Близко к этому примыкают подходы, в которых гиперболическая система интегрируется вдоль направления распространения колебаний. Например, такой подход развивали Доброхотов, Назайкинский, Шафаревич, Секерж-Зенькович, Аникин, Толченников и другие (см., например, [320, 321] и цитированную литературу). В указанных работах задача решалась в два этапа: сначала на основе вариационного принципа вычислялись лучевые траектории, затем вдоль них проводились одномерные расчеты фронта волны.

Эти авторы применяли данный подход к расчетам коротковолновых радиотрасс в ионосфере, моделированию распространения океанских волн и формирования цунами и некоторым другим задачам. Во всех этих задачах свойства среды, в которой распространяется волна, плавно зависят от координаты, т.е. границы раздела отсутствуют.

Аналогичный подход развивали Форбс и Алонсо применительно к задачам волновой оптики (дифракции, распространения электромагнитных полей по волноводам и др.) [322–329]. Эти авторы рассматривали отражение и преломление на одной границе раздела. Они вводили лучи для отраженной и преломленной волн. Однако при наличии нескольких границ раздела возникают множественные переотражения, причем в прозрачной среде таких перетражений бесконечно много. Обобщение метода Форбса-Алонсо на этот случай сталкивается с серьезными трудностями. Суммарное поле содержит бесконечное число слагаемых. Обрезание такого ряда вносит погрешность, величина которой требует дополнительных исследований.

Тем не менее, описанные полуаналитические методы намного экономичнее прямого численного моделирования многомерной задачи, поэтому он представляется перспективными.

Задачи п. 7.1.1 – 7.1.4 являются двумерными. В данной работе предлагается подход, позволяющий свести их к одномерной постановке. Он применим как в среде с плавно меняющимся показателем преломления, так и в слоистой (имеющей несколько границ раздела). Этим предлагаемый подход отличается от методов группы Доброхотова и метода Форбса-Алонсо. Он был назван ***методом оптических путей***. Он сводится к интегрированию уравнений Максвелла вдоль направления распространения (лучевой траектории) падающей и преломленной волн.

7.4.2. Лучевые траектории

1° Рассмотрим стационарные задачи п. 7.1.1 и 7.1.2. Пусть сначала объемные и поверхностные токи отсутствуют $\mathbf{J} = 0$, $\mathbf{j} = 0$. Обобщение на случай проводящих пластин будет построено далее (см. п. 7.4.8). Материал пластин может быть как однородным (см. задачу п. 7.1.2), так и неоднородным (см. задачу п. 7.1.1).

2° На каждой q-й границе раздела падающая волна частично отражается и, преломляясь, частично проходит далее. У падающей и прошедшей волны z-компонента волнового вектора положительна $k_z > 0$. Такие волны будем называть прямыми. У отраженной волны z-компонента волнового вектора отрицательна $k_z > 0$. Такие волны будем называть обратными. Прямая прошедшая волна падает на следующую $(q+1)$-ю границу раздела, причем угол падения равен углу преломления для q-й границы раздела. На $(q+1)$-й границе раздела волна претерпевает отражение и преломление. Отраженная обратная волна возвращается к q-й границе и также испытывает на ней отражение и преломление. Волна, отраженная от q-й границы, становится прямой и вместе с другими прямыми волнами падает на $(q+1)$-ю границу.

Число таких переотражений очень велико (в случае пренебрежимо малого поглощения переотражений бесконечно много). При этом в каждой пластине все прямые волны имеют одни и те же углы падения и преломления, все обратные волны имеют одни и те же углы отражения. Поэтому можно ввести единую лучевую траекторию для прямых волн.

2° Лучевую траекторию построим в рамках геометрической оптики. Для этого необходимо использовать принцип Ферма (аналогично [321]). Он применим, поскольку предполагается, что пространственная дисперсия отсутствует. Принцип Ферма приводит к задаче на экстремум функционала времени распространения света в среде

$$\int_0^a \frac{\mathrm{n}(z)}{c}\sqrt{1+(x'(z))^2}\,dz \to \min\,. \tag{7.11}$$

Здесь $c/\mathrm{n}(z)$ – скорость распространения света в среде. Это уравнение определяет лучевую траекторию $x(z)$ падающей волны. В неоднородной среде данная лучевая траектория является криволинейной. Лучевая траектория отраженной волны зеркально симметрична траектории падающей волны относительно плоскости Oyz. Чтобы построить ее, нужно сделать замену $x \to -x$ при сохранении знака z.

Методы решения таких задач обсуждаются в [330–332]. Возможно также применение прямого сеточного метода [12].

3° В простейшем случае, если рассеиватель составлен из однородных изотропных диэлектрических пластин, эта задача допускает несложное аналитическое решение. Прямые и обратные волны распространяются по прямым линиям, направление которых в q-й пластине определяется законом Снеллиуса (который, как известно, есть следствие принципа Ферма). Угол преломления β_q в q-й границе раздела определяется равенством

$$\beta_q = \arcsin\frac{\mathrm{n}_{q+1}}{\mathrm{n}_q}\sin\alpha_q, \tag{7.12}$$

где α_q – угол падения на q-ю границу раздела.

Пример лучевой траектории для одной границы раздела между однородными средами приведен на рис. 7.1, 7.2. Эта лучевая траектория является негладкой: она испытывает излом на каждой границе раздела. Обозначим координату вдоль лучевой траектории через η. Для случая рис. 7.1, 7.2 преобразование координат $z \to \eta$ выполняется по следующему правилу:

$$\begin{aligned} \eta &= \frac{z}{\cos\alpha}, \quad z < 0, \\ \eta &= \frac{z}{\cos\beta}, \quad z > 0. \end{aligned} \tag{7.13}$$

Если границ раздела несколько, то преобразование (7.13) обобщается очевидным образом.

4° В качестве пространственной координаты выберем ***лучевую траекторию*** прямой волны.

7.4.3. Неизвестные функции

1° Рассмотрим падающую волну. Как в однородной, так и в неоднородной среде на лучевой траектории поля $\mathbf{E}$ и $\mathbf{H}$ ортогональны $\mathbf{k}$ и имеют только одну компоненту, равную комплексной амплитуде соответствующего вектора. Поэтому задача, в которой присутствует только падающая волна, является одномерной. То же справедливо для отраженной волны.

2° Во всех пластинах рассеивателя, кроме последней $\xi_{Q-1} \leqslant z \leqslant \xi_Q$, поле является суперпозицией падающей и отраженной волн. В этом случае суммарные векторы $\mathbf{E}$ и $\mathbf{H}$ не ортогональны $\mathbf{k}$. Однако, согласно закону отражения, углы между полевыми векторами падающей и отраженной волн и осями координат одинаковы. Так, для s-поляризации угол между полем $\mathbf{H}_{\text{inc}}$ падающей волны и осью z равен углу между полем $\mathbf{H}_{\text{refl}}$ отраженной волны и осью z (аналогично для оси x). То же справедливо для полей $\mathbf{E}_{\text{inc}}$, $\mathbf{E}_{\text{refl}}$ в случае p-поляризации. Поэтому проекции суммарного поля на оси координат вычисляются следующим образом:

$$\begin{aligned} H_z &= (H_{\text{inc}})_z + (H_{\text{refl}})_z = (H_{\text{inc}} + H_{\text{refl}}) \sin\alpha, \quad \text{для } s\text{-поляризации}, \\ E_z &= (E_{\text{inc}})_z + (E_{\text{refl}})_z = (E_{\text{inc}} + E_{\text{refl}}) \sin\alpha, \quad \text{для } p\text{-поляризации}. \end{aligned} \tag{7.14}$$

Выражения для E_x, H_x получаются из (7.14) заменой $\sin \to \cos$.

3° Таким образом, как и в случае нормального падения, решение полностью определяется ***суммой комплексных амплитуд*** падающей и отраженной волн. Поэтому для расчета можно использовать одномерную схему, в которой неизвестной функцией является указанная сумма.

7.4.4. Условия сопряжения

Интегральные уравнения Максвелла в изотропной среде инвариантны относительно поворота системы координат. Поэтому, чтобы перейти к координате вдоль лучевой траектории, достаточно модифицировать условия сопряжения на границе раздела сред.

Рассмотрим падение волны на одну границу раздела (см. рис. 7.1, 7.2). Условия сопряжения для тангенциальных компонент полей на этой границе имеют вид

$$(E_\tau)_1 = (E_\tau)_2,\ (H_\tau)_1 = (H_\tau)_2. \tag{7.15}$$

Здесь среда 1 расположена до границы раздела, а среда 2 – после. Для s-поляризации условия сопряжения имеют вид

$$E_1 = E_2,\ H_1 \cos\alpha = H_2 \cos\beta, \tag{7.16}$$

где α – заданный угол падения, β – угол преломления. Для p-поляризации условия сопряжения записываются следующим образом:

$$H_1 = H_2,\ E_1 \cos\alpha = E_2 \cos\beta. \tag{7.17}$$

Условия сопряжения непосредственно входят в разностную схему. По сравнению со случаем нормального падения добавляются только множители $\cos\alpha$ и $\cos\beta$. Если границ раздела несколько, то условия (7.16), (7.17) записываются на каждой из них.

Чтобы учесть полное внутреннее отражение, введем чисто мнимое волновое число $k \to ik$, если $(\mathrm{Re}\,\mathrm{n}_1/\mathrm{Re}\,\mathrm{n}_2)\sin\alpha > 1$. Такая замена справедлива, если среда не является генерирующей.

7.4.5. Эффективная толщина

После перехода к координате вдоль лучевой траектории эффективная толщина пластин отличается от физической толщины. Потребуем, чтобы пластины эффективной толщины обеспечивала правильный набег фазы.

Рассмотрим наклонное падение плоской волны на плоско-параллельную пластину (интерферометр Фабри-Перо). Как известно, разность фаз у волны, однократно прошедшей туда и обратно через пластинку, и волны, отразившейся от наружной поверхности пластины, равна

$$\delta = \frac{2\pi}{\lambda} 2h\mathrm{n}\cos\beta. \tag{7.18}$$

Чтобы обеспечить такую же разность фаз при движении вдоль лучевой траектории, заменим физическую толщину на эффективную

$$h \to h\cos\beta. \tag{7.19}$$

Если в среде имеется поглощение, то есть $\operatorname{Im} \mathrm{n} \neq 0$, то угол β формально оказывается комплексным. При этом набег фазы δ также является комплексным, то есть учитывает затухание волны. В этом случае в (7.19) нужно взять $|\cos\beta|$.

Такая пластинка эффективной толщины будет давать такой же спектр отражения, что исходная пластинка при наклонном падении. При наличии нескольких пластин толщину каждой заменим на эффективную.

Замечание 1. Для прозрачной среды без поглощения замена пластинки на эффективную является точной. Если поглощением пренебречь нельзя, то метод оптических путей вносит некоторую физическую погрешность. Величина этой погрешности проиллюстрирована в п. 7.5.5.

Замечание 2. Строго говоря, в диспергирующей среде эффективные толщины для различных частот оказываются разными. Если среда является еще и неоднородной, то для волн с различными частотами будут несколько отличаться лучевые траектории.

Однако для реальных диэлектрических материалов в оптическом диапазоне это различие невелико. Поэтому мы им пренебрегаем и вычисляем лучевые траектории и эффективные толщины слоев по частоте, соответствующей середине рассматриваемого диапазона. Далее на конкретных примерах (см. п. 7.5.6) будет показано, что вносимая при этом погрешность не превышает 0.1%, что гарантированно перекрывает потребности практики в задачах фотоники и плазмоники.

7.4.6. Разностная схема

Рассмотрим оптически эквивалентный рассеиватель. Эффективные толщины пластин равны $(\xi_{q+1} - \xi_q)\cos\beta_q$, $1 \leqslant q \leqslant Q-1$, где $\xi_0 = 0$. Введем специаль-

ную сетку, в которой указанные точки являются узлами; остальные узлы могут быть расставлены произвольно. Разностная схема для случая s-поляризации будет иметь следующий вид:

$$H_{2n-1} - H_{2n-2} = \frac{i\omega}{2c}\varepsilon_{n-1/2}\Delta z_{n-1/2}(E_{2n-1} + E_{2n-2}), \quad 1 \le n \le N, \tag{7.20}$$

$$E_{2n-1} - E_{2n-2} = \frac{i\omega}{2c}\mu_{n-1/2}\Delta z_{n-1/2}(H_{2n-1} + H_{2n-2}), \quad 1 \le n \le N, \tag{7.21}$$

$$E_{2n} = E_{2n-1}, \quad H_{2n}\cos\beta_n = H_{2n-1}\cos\alpha_n, \quad 1 \le n \le N-1, \tag{7.22}$$

$$\begin{gathered}\sqrt{\varepsilon_0}E_0 + \sqrt{\mu_0}H_0 = 2\sqrt{\varepsilon_0}E^0, \ z = 0 \\ \sqrt{\varepsilon_N}E_{2N-1} - \sqrt{\mu_N}H_{2N-1} = 2\sqrt{\varepsilon_N}E^a, \ z = a.\end{gathered} \tag{7.23}$$

Здесь α_0 и β_0 – угол падения и угол преломления на границе расчетной области, α_1 и β_1 – угол падения и угол преломления, соответствующие первому узлу сетки и т.д. При этом β_0 вычисляем по заданному α_0 по формуле (7.12), затем полагаем $\alpha_1 = \beta_0$, β_1 определяем по α_1 согласно (7.12) и т.д. Если в n-м узле граница раздела отсутствует, то преломления не происходит, и $\beta_n = \alpha_n$ по построению.

В случае p-поляризации нужно заменить равенства (7.22) на

$$E_{2n}\cos\beta_n = E_{2n-1}\cos\alpha_n, \quad H_{2n} = H_{2n-1}, \quad 1 \le n \le N-1. \tag{7.24}$$

Решение разностной схемы (7.20) – (7.24) соответствует фиксированной координате x. Чтобы получить решение исходной двумерной задачи, необходимо сделать обратную замену переменных $\eta \to z$ по формуле (7.13) и умножить решение на

$$\exp\left(ix_m\frac{\omega}{c}\sqrt{\varepsilon_n\mu_n}\sin\alpha_n\right). \tag{7.25}$$

Здесь $\{x_m\}$ – сетка по координате x. Значения материальных параметров ε_n, μ_n целесообразно вычислять в узлах сетки. Множитель (7.25) описывает распространение волны вдоль границ раздела.

Напомним, что в (7.20) – (7.24) неизвестными являются не проекции, а комплексные амплитуды полевых векторов. Заметим, что они претерпевают разрывы на каждой границе раздела сред. Однако бикомпактные схемы, постро-

енные в главе 5, позволяют легко преодолеть эту трудность. При этом метод оптических путей существенно расширяет область применимости одномерных бикомпактных схем.

Сходимость схемы (7.20) – (7.24) обосновывается полностью аналогично п. 5.4. В частности, справедливы те же оценки устойчивости по граничным условиям.

7.4.7. Клиновидная пластина

Предложенный метод можно обобщить на случай клиновидных пластин, когда границы раздела являются плоскими, но не параллельными. Построим такое обобщение, следуя [264].

Набег фазы волны, прошедшей туда и обратно через пластинку, приближенно описывается формулой (7.18), где h – толщина пластины в том месте, где происходит отражение света. Так, если угол при вершине клина равен γ, то на расстоянии x от вершины толщина пластинки равна $h(x) = x \operatorname{tg} \gamma$. Поэтому для клиновидной пластины эффективная толщина вводится по формуле (7.19), в которой нужно подставить «местную» физическую толщину $h(x)$.

В условиях сопряжения необходимо учитывать, что углы α и β отсчитываются от нормали к границам раздела.

Сеточная задача (7.20) – (7.24) решается при каждом фиксированном x_l. Решение умножается на множитель (7.25), где вместо x_m нужно подставить $x_m - x_l$ (то есть на то расстояние, «проходит» волна от источника x_l до точки наблюдения x_m). Затем полученные решения необходимо просуммировать с весом, обратным числу шагов сетки по x.

7.4.8. Индуцированные токи

S-поляризация. В этом случае векторы электрического поля $\mathbf{E}_{\text{inc}} = \{0, E_{\text{inc}}, 0\}$ в падающей и $\mathbf{E}_{\text{refl}} = \{0, E_{\text{refl}}, 0\}$ направлены вдоль оси y. Поэтому вектор $\mathbf{J}^{\text{ind}} = \sigma \mathbf{E}$ также направлен вдоль той же оси. Эти токи излучают электро-

магнитные волны, в которых вектор $\mathbf{E}_{\text{emit}} = \{0, E_{\text{emit}}, 0\}$ направлен так же, как векторы $\mathbf{E}_{\text{inc}}$ и $\mathbf{E}_{\text{refl}}$. Амплитуда суммарного поля есть сумма амплитуд полей падающей, отраженной и переизлученной волн. Поэтому схема для случая s-поляризованных волн в проводящей среде будет иметь следующий вид:

$$\begin{gathered} H_{2n-1} - H_{2n-2} - \\ -(i\omega c^{-1}\varepsilon_{n-1/2} - 4\pi c^{-1}\sigma_{n-1/2})\Delta z_{n-1/2}(E_{2n-1} + E_{2n-2}) = 0, \\ 1 \le n \le N, \end{gathered} \tag{7.26}$$

$$\begin{gathered} E_{2n-1} - E_{2n-2} - i\omega(2c)^{-1}\mu_{n-1/2}\Delta z_{n-1/2}(H_{2n-1} + H_{2n-2}) = 0, \\ 1 \le n \le N, \end{gathered} \tag{7.27}$$

$$E_{2n} = E_{2n-1}, \quad H_{2n}\cos\beta_n = H_{2n-1}\cos\alpha_n, \quad 1 \le n \le N-1, \tag{7.28}$$

$$\begin{gathered} \sqrt{\varepsilon_0}E_0 + \sqrt{\mu_0}H_0 = 2\sqrt{\varepsilon_0}E^0, \ z = 0 \\ \sqrt{\varepsilon_N}E_{2N-1} - \sqrt{\mu_N}H_{2N-1} = 2\sqrt{\varepsilon_N}E^a, \ z = a. \end{gathered} \tag{7.29}$$

***P*-поляризация.** При p-поляризации векторы $\mathbf{E}_{\text{inc}}$ падающей и $\mathbf{E}_{\text{refl}}$ отраженной волн не лежат в плоскости границ раздела: они имеют x- и z-компоненты. Объемные токи $\mathbf{J}^{\text{ind}}$ направлены так же, как вектор $\mathbf{E}_{\text{inc}} + \mathbf{E}_{\text{refl}}$. Поэтому в излучаемой ими волне вектор $\mathbf{E}_{\text{emit}}$ параллелен сумме электрических полей падающей и отраженной волн $\mathbf{E}_{\text{inc}} + \mathbf{E}_{\text{refl}}$.

Таким образом, разностная схема для случая p-поляризованной волны в проводящей среде имеет следующий вид:

$$\begin{gathered} H_{2n-1} - H_{2n-2} - \\ -(i\omega c^{-1}\varepsilon_{n-1/2} - 4\pi c^{-1}\sigma_{n-1/2})\Delta z_{n-1/2}(E_{2n-1} + E_{2n-2}) = 0, \\ 1 \le n \le N, \end{gathered} \tag{7.30}$$

$$\begin{gathered} E_{2n-1} - E_{2n-2} - i\omega(2c)^{-1}\mu_{n-1/2}\Delta z_{n-1/2}(H_{2n-1} + H_{2n-2}) = 0, \\ 1 \le n \le N, \end{gathered} \tag{7.31}$$

$$E_{2n}\cos\beta_n = E_{2n-1}\cos\alpha_n, \ H_{2n} = H_{2n-1}, \ 1 \le n \le N-1 \tag{7.32}$$

$$\begin{gathered} \sqrt{\varepsilon_0}E_0 + \sqrt{\mu_0}H_0 = 2\sqrt{\varepsilon_0}E^0, \ z = 0 \\ \sqrt{\varepsilon_N}E_{2N-1} - \sqrt{\mu_N}H_{2N-1} = 2\sqrt{\varepsilon_N}E^a, \ z = a. \end{gathered} \tag{7.33}$$

7.4.9. Нестационарные задачи

Нестационарные задачи п. 7.1.3, 7.1.4 методом спектрального разложения сводятся к набору стационарных задач п. 7.1.1, 7.1.2. Для каждой них записываются разностные схемы (7.20) – (7.24), (7.26) – (7.29) либо (7.30) – (7.33). Решим их и выполним обратное преобразование Фурье. Это даст искомое решение нестационарных задач п. 7.1.3, 7.1.4.

7.4.10. Сравнение с аналогичными подходами

Сравним область применимости метода оптических путей и известных подходов.

Как отмечалось выше, методы группы Доброхотова применимы к задачам, в которых свойства среды плавно зависят от координаты, т.е. границы раздела отсутствуют. Метод Форбса-Алонсо разработан для сред с плавно меняющимся показателем преломления. Обобщение этих подходов на задачи в слоистых средах с несколькими границами раздела сталкивается со значительным трудностями из-за множественных переотражений. Предлагаемый метод оптических путей построен как для градиентных сред с плавно меняющимся показателем преломления, так и для слоистых.

В оптических задачах широко применяются методы матриц рассеяния. Однако они применимы для кусочно-однородных сред: внутри слоев показатели преломления не должны зависеть от координаты. Метод оптических путей применим к задачам, в которых пластины могут быть пространственно неоднородными.

Методы матриц рассеяния и метод Форбса-Алонсо применимы только для стационарных задач. Метод оптических путей построен как для стационарных, так и для нестационарных задач.

Таким образом, в рамках постановок п. 7.1.1 – 7.1.4 метод оптических путей применим к более широкому кругу задач.

7.4.11. Пакет программ

Предлагаемый подход (метод оптических путей, метод спектрального разложения и бикомпактные схемы для стационарных подзадач) реализован в виде прикладного пакета программ `BiSpDec` в среде Matlab. В нем реализовано решение системы уравнений Максвелла при наклонном падении плоской линейно поляризованной волны на набор плоско-параллельных диэлектрических пластин. Программа вычисляет электромагнитное поле, спектр отражения от рассеивателя, кросс-корреляционную функцию и время жизни связанного состояния с апостериорной асимптотически точной оценки погрешности. Этот пакет непосредственно использовался при приведении расчетов п. 7.6. Пакет распространяется по свободной лицензии BSD-3-clause и доступен по ссылке `https://github.com/ABelov91/BiSpDec/`.

7.5. Верификация

Проведем верификацию предложенного метода на тестовых задачах с известными точными решениями. В задачах этого раздела все размерные величины нормированы (то есть являются относительными).

7.5.1. Граница раздела, s-поляризация

1° Рассмотрим задачу о наклонном падении плоской монохроматической волны на плоскую границу раздела. Пусть последняя соответствует $z = b > 0$. Пусть при $0 < z < 0$ материальные параметры среды есть ε_1, μ_1; при $b < z < a - \varepsilon_2$, μ_2. Пусть амплитуда электрического поля падающей волны равна $\mathbf{E}^0$. В случае s-поляризации вектор $\mathbf{E}$ перпендикулярен плоскости падения (см. рис. 7.1).

2° Приведем математическую постановку этой задачи. Выделим участок границы раздела, ограниченный плоскостями $x = 0$ и $x = d > 0$. В области

$0 \le x \le d$, $0 \le z \le a$ электромагнитные поля удовлетворяют уравнениям Максвелла (5.4). На границе раздела $z = b$ необходимо поставить условия сопряжения (5.5). На границах области $z = 0$, $z = a$, $x = 0$, $x = d$ запишем двумерные условия излучения (7.2) – (7.5).

3° Построим точное решение этой задачи. Очевидно, оно равно сумме падающей и отраженной волн при $z < b$ и прошедшую волна при $z > b$. Амплитудные коэффициенты отраженной и прошедшей волн выражаются общеизвестными формулами Френеля. Таким образом, при $0 < z < b$ решение имеет вид

$$\begin{gathered}\mathbf{E} = \mathbf{e}_y \exp(ix\tilde{k}_1 \sin\alpha)\left[E^0 \exp(iz\tilde{k}_1\cos\alpha) + C_2\exp(-iz\tilde{k}_1\cos\alpha)\right], \\ \mathbf{H} = (\mathbf{e}_x\cos\alpha + \mathbf{e}_z\sin\alpha)\exp(ix\tilde{k}_1\sin\alpha)\sqrt{\frac{\varepsilon_1}{\mu_1}}\cdot \\ \cdot\left[E^0\exp(iz\tilde{k}_1\cos\alpha) - C_2\exp(-iz\tilde{k}_1\cos\alpha)\right],\end{gathered} \tag{7.34}$$

$$C_2 = E^0\frac{\sqrt{\varepsilon_1/\mu_1}\cos\alpha - \sqrt{\varepsilon_2/\mu_2}\cos\beta}{\sqrt{\varepsilon_1/\mu_1}\cos\alpha + \sqrt{\varepsilon_2/\mu_2}\cos\beta}e^{i\tilde{k}_1 b}, \quad \tilde{k}_1 = \frac{\omega}{c}\sqrt{\varepsilon_1\mu_1}. \tag{7.35}$$

При $b < z \leqslant a$ оно равно

$$\begin{gathered}\mathbf{E} = \mathbf{e}_y C_3\exp(ix\tilde{k}_2\sin\alpha + iz\tilde{k}_2\cos\alpha), \\ \mathbf{H} = (\mathbf{e}_x\cos\beta + \mathbf{e}_z\sin\beta)C_3\sqrt{\frac{\varepsilon_2}{\mu_2}}\exp(ix\tilde{k}_2\sin\alpha + iz\tilde{k}_2\cos\alpha),\end{gathered} \tag{7.36}$$

$$C_3 = E^0\frac{2\sqrt{\varepsilon_1/\mu_1}\cos\alpha}{\sqrt{\varepsilon_1/\mu_1}\cos\alpha + \sqrt{\varepsilon_2/\mu_2}\cos\beta}e^{i(\tilde{k}_1-\tilde{k}_2)b}, \quad \tilde{k}_2 = \frac{\omega}{c}\sqrt{\varepsilon_2\mu_2}. \tag{7.37}$$

4° Построим теперь решение задачи в переменной η, то есть вдоль лучевой траектории. Постановка этой задачи имеет вид

$$\frac{\partial E}{\partial\eta} = \frac{i\mu_1\omega}{c}H; \quad \frac{\partial H}{\partial\eta} = \frac{i\varepsilon_1\omega}{c}E, \quad 0 \leqslant \eta \leqslant b\cos\alpha; \tag{7.38}$$

$$\frac{\partial E}{\partial\eta} = \frac{i\mu_2\omega}{c}H; \quad \frac{\partial H}{\partial\eta} = \frac{i\varepsilon_2\omega}{c}E, \quad b\cos\alpha \leqslant \eta \leqslant a\cos\beta; \tag{7.39}$$

$$\frac{\partial E}{\partial\eta} + \frac{i\omega}{c}E = 2i\tilde{k}E^0,\ \eta = 0; \quad \frac{\partial E}{\partial\eta} - \frac{i\omega}{c}E = 0,\ \eta = a\cos\beta. \tag{7.40}$$

$$E|_{\eta=b\cos\alpha-0} = E|_{\eta=b\cos\alpha+0}, \quad H\cos\alpha|_{\eta=b\cos\alpha-0} = H\cos\beta|_{\eta=b\cos\alpha+0}. \tag{7.41}$$

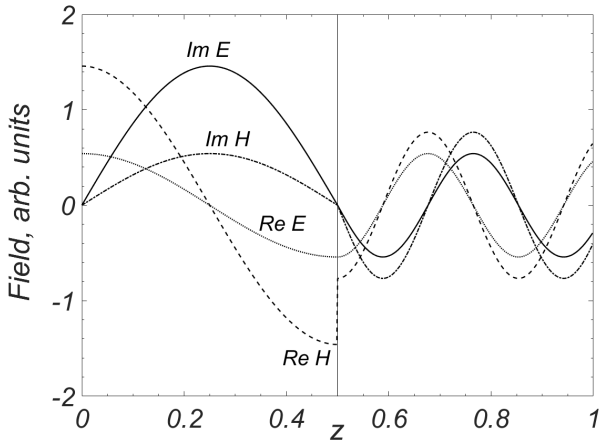

Рис. 7.3. Амплитуды полей в задаче о падении s-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

Решение строится аналогично (6.10) – (6.13). Оно имеет вид

$$E = E^0 e^{i\tilde{k}_1\eta} + C_2 e^{-i\tilde{k}_1\eta}, \quad H = \sqrt{\frac{\varepsilon_1}{\mu_1}}(E^0 e^{i\tilde{k}_1\eta} - C_2 e^{-i\tilde{k}_1\eta}), \quad 0 \leqslant \eta \leqslant b\cos\alpha; \tag{7.42}$$

$$E = C_3 e^{i\tilde{k}_2\eta}, \quad H = \sqrt{\frac{\varepsilon_2}{\mu_2}} C_3 e^{i\tilde{k}_2\eta}, \quad b\cos\alpha \leqslant z \leqslant a\cos\beta. \tag{7.43}$$

Здесь коэффициенты C_2, C_3 определяются выражениями (7.35), (7.37).

Легко видеть, что решение (7.34) – (7.37) получается из (7.42), (7.43) домножением на $\exp(ix\tilde{k}_1\sin\alpha)$ либо $\exp(ix\tilde{k}_2\sin\beta)$, что соответствует распространению волны вдоль оси x, и на $\cos\alpha$ либо $\sin\alpha$ в случае проекций на оси x и z соответственно (ср. с формулой (7.14)). Таким образом, два точных решения, полученных разными способами, согласуются между собой. Это подтверждает правильность метода оптических путей.

5° Положим $a = 1$, $b = 0.5$, $\omega = 2\pi$, $\alpha = \pi/3$, $\varepsilon_1 = \mu_1 = 1$, $\varepsilon_2 = 4$, $\mu_2 = 2$, $c = 1$, $E^0 = 1$, $E^a = 0$. На трех следующих рисунках показано решение, полученное методом оптических путей. На рис. 7.3 показаны амплитуды полей, то есть непосредственно те функции, которые вычисляются в ходе сеточного расчета. Компонента $\mathrm{Re}\, E$ является непрерывной и гладкой. Компоненты $\mathrm{Im}\, E$, $\mathrm{Im}\, H$ являются непрерывными, но негладкими: на границе раздела они имеют слабый разрыв. Компонента $\mathrm{Re}\, H$ на границе раздела испытывает сильный разрыв.

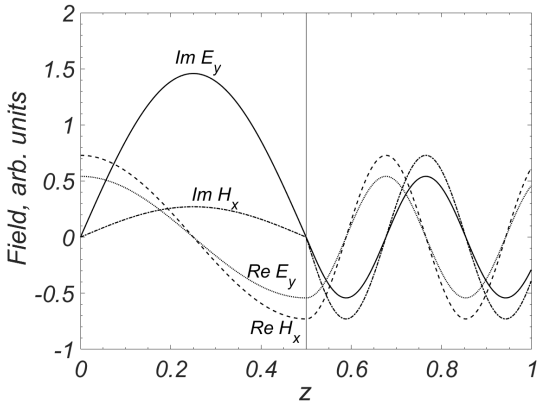

Рис. 7.4. Тангенциальные компоненты полей в задаче о падении s-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

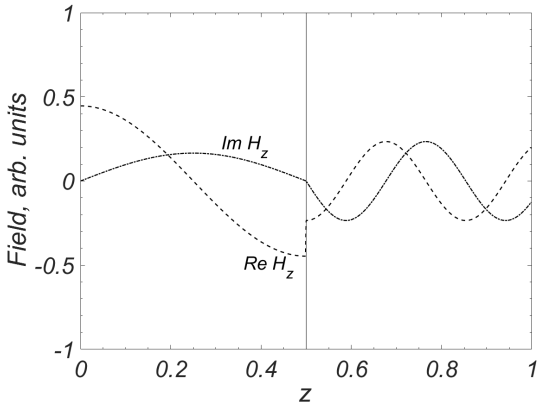

Рис. 7.5. Нормальная компонента поля H в задаче о падении s-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

На рис. 7.4 представлены тангенциальные компоненты векторов E_y, H_x. Видно, что все они непрерывны. При этом E_y совпадает с рис. 7.3. Компоненты $\mathrm{Re}\, E_y$ и $\mathrm{Re}\, H_x$ являются гладкими, а $\mathrm{Im}\, E_y$ и $\mathrm{Im}\, H_x$ испытывают излом на границе раздела сред.

На рис. 7.5 приведена нормальная компонента H_z. Видно, что компонента $\mathrm{Im}\, H_z$ непрерывна, но испытывает излом на границе раздела. Компонента $\mathrm{Re}\, H_z$ претерпевает там разрыв.

Расчет проводился на наборе сгущающихся сеток. Погрешность определялась двумя способами: (а) при помощи непосредственного сравнения численного решения с точным (7.42), (7.43) и (б) апостериорно по методу Ричардсона-Калитина. На рис. 7.6 представлены погрешности, полученные в данном расчете. График дан в двойном логарифмическом масштабе, поэтому прямая линия соответствует степенной сходимости, причем наклон прямой равен порядку точности схемы. Видно, что линии погрешности стремятся к прямым с наклоном -2. Видно также, что на прямолинейном участке кривых погрешности, вычисленные по методу сгущения сеток, практически совпадают с погрешностями

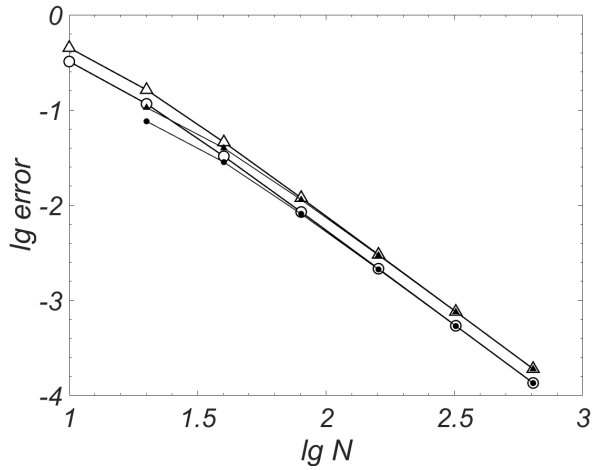

Рис. 7.6. Погрешность решения в задаче о падении s-поляризованной волны на диэлектрическую границу раздела: $\circ$ – E_x, $\triangle$ – H_y, светлые маркеры – разность численного и точного решений, темные маркеры – оценки по методу Ричардсона.

относительно точного решения. Это убедительно подтверждает работоспособность предложенного метода.

Из приведенных расчетов видно, что данная задача очень сложна для сеточных методов. Тем не менее, бикомпактная схема на специальных сетках успешно с ней справляется.

7.5.2. Граница раздела, p-поляризация

1° Рассмотрим предыдущую задачу в случае p-поляризации (см. рис. 7.2). Теперь вектор $\mathbf{H}$ перпендикулярен плоскости падения.

2° Точное решение строится аналогично случаю s-поляризации. Для этого достаточно в (7.34), (7.36) поменять местами поля $\mathbf{E}$ и $\mathbf{H}$ (при этом отраженная волна для поля $\mathbf{E}$ должна быть со знаком «плюс», для поля $\mathbf{H}$ – со знаком «минус») и заменить C_2, C_3 на коэффициенты Френеля для p-поляризации.

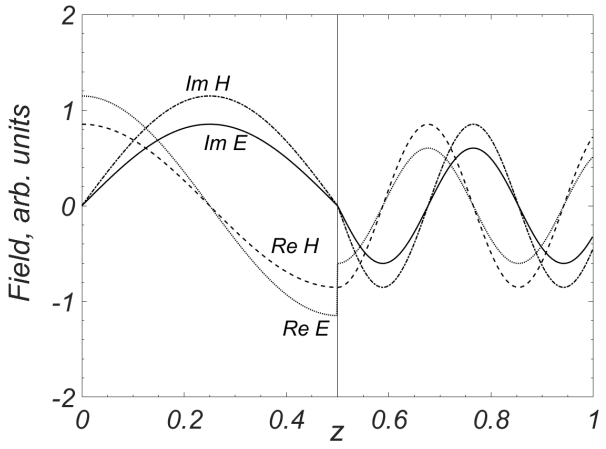

Рис. 7.7. Амплитуды полей в задаче о падении p-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

При $0 \leqslant z \leqslant b$ решение равно

$$\begin{gathered}\mathbf{H} = \mathbf{e}_y \exp(ix\tilde{k}_1 \sin\alpha)\cdot \\ \cdot\sqrt{\frac{\varepsilon_1}{\mu_1}}\left[E^0 \exp(iz\tilde{k}_1\cos\alpha) - C_2\exp(-iz\tilde{k}_1\cos\alpha)\right], \\ \mathbf{E} = (\mathbf{e}_x\cos\alpha + \mathbf{e}_z\sin\alpha)\exp(ix\tilde{k}_1\sin\alpha)\cdot \\ \cdot\left[E^0\exp(iz\tilde{k}_1\cos\alpha) + C_2\exp(-iz\tilde{k}_1\cos\alpha)\right],\end{gathered} \tag{7.44}$$

$$C_2 = E^0\frac{\sqrt{\varepsilon_1/\mu_1}\cos\beta - \sqrt{\varepsilon_2/\mu_2}\cos\alpha}{\sqrt{\varepsilon_1/\mu_1}\cos\beta + \sqrt{\varepsilon_2/\mu_2}\cos\alpha}e^{i\tilde{k}_1 b}, \quad \tilde{k}_1 = \frac{\omega}{c}\sqrt{\varepsilon_1\mu_1}. \tag{7.45}$$

При $b < z \leqslant a$ оно равно

$$\begin{gathered}\mathbf{H} = \mathbf{e}_y C_3\sqrt{\frac{\varepsilon_2}{\mu_2}}\exp(ix\tilde{k}_2\sin\alpha + iz\tilde{k}_2\cos\alpha), \\ \mathbf{E} = (\mathbf{e}_x\cos\beta + \mathbf{e}_z\sin\beta)C_3\exp(ix\tilde{k}_2\sin\alpha + iz\tilde{k}_2\cos\alpha),\end{gathered} \tag{7.46}$$

$$C_3 = E^0\frac{2\sqrt{\varepsilon_1/\mu_1}\cos\alpha}{\sqrt{\varepsilon_1/\mu_1}\cos\beta + \sqrt{\varepsilon_2/\mu_2}\cos\alpha}e^{i(\tilde{k}_1-\tilde{k}_2)b}, \quad \tilde{k}_2 = \frac{\omega}{c}\sqrt{\varepsilon_2\mu_2}. \tag{7.47}$$

3° Построим теперь решение задачи вдоль лучевой траектории. Постановка включает в себя уравнения (7.38) – (7.40) и условие сопряжения

$$H|_{\eta=b\cos\alpha-0} = H|_{\eta=b\cos\alpha+0}, \quad E\cos\alpha|_{\eta=b\cos\alpha-0} = E\cos\beta|_{\eta=b\cos\alpha+0}. \tag{7.48}$$

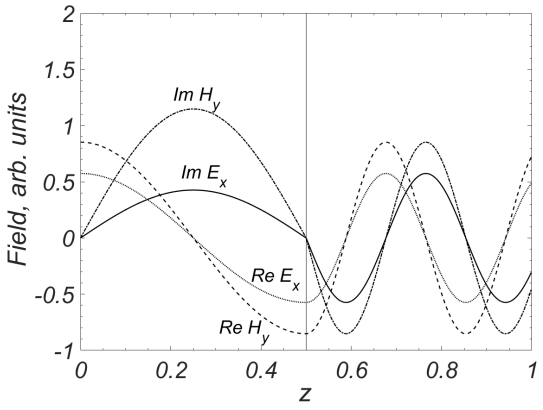

Рис. 7.8. Тангенциальные компоненты полей в задаче о падении p-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

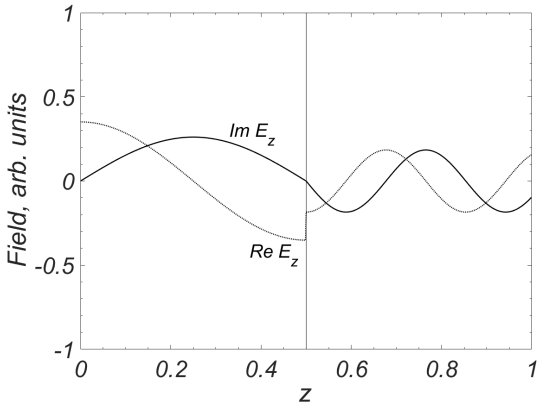

Рис. 7.9. Нормальная компонента поля E в задаче о падении p-поляризованной волны на диэлектрическую границу раздела. Вертикальная прямая – граница раздела.

Решение этой задачи имеет вид

$$E = E^0 e^{i\tilde{k}_1\eta} + C_2 e^{-i\tilde{k}_1\eta}, \quad H = \sqrt{\frac{\varepsilon_1}{\mu_1}}(E^0 e^{i\tilde{k}_1\eta} - C_2 e^{-i\tilde{k}_1\eta}), \qquad 0 \leqslant \eta \leqslant b\cos\alpha; \tag{7.49}$$

$$E = C_3 e^{i\tilde{k}_2\eta}, \quad H = \sqrt{\frac{\varepsilon_2}{\mu_2}} C_3 e^{i\tilde{k}_2\eta}, \quad b\cos\alpha \leqslant z \leqslant a\cos\beta. \tag{7.50}$$

Здесь коэффициенты C_2, C_3 определяются выражениями (7.45), (7.47).

4° Зададим такие же параметры, как в предыдущей задаче: $a = 1$, $b = 0.5$, $\omega = 2\pi$, $\alpha = \pi/3$, $\varepsilon_1 = \mu_1 = 1$, $\varepsilon_2 = 4$, $\mu_2 = 2$, $c = 1$, $E^0 = 1$, $E^a = 0$.

На рис. 7.7 показаны амплитуды полей. Компонента $\operatorname{Re} H$ является непрерывной и гладкой. Компоненты $\operatorname{Im} E$, $\operatorname{Im} H$ непрерывны, но претерпевают излом на границе раздела. Компонента $\operatorname{Re} E$ испытывает сильный разрыв.

На рис. 7.8 представлены тангенциальные компоненты векторов E_x, H_y. Видно, что они непрерывны. При этом H_y совпадает с рис. 7.7. Компоненты $\operatorname{Re} E_x$ и $\operatorname{Re} H_y$ являются гладкими, а $\operatorname{Im} E_x$ и $\operatorname{Im} H_y$ испытывают излом на границе раздела сред.

На рис. 7.9 приведена нормальная компонента E_z. Видно, что $\operatorname{Im} E_z$ имеет слабый разрыв на границе раздела, а $\operatorname{Re} E_z$ – сильный.

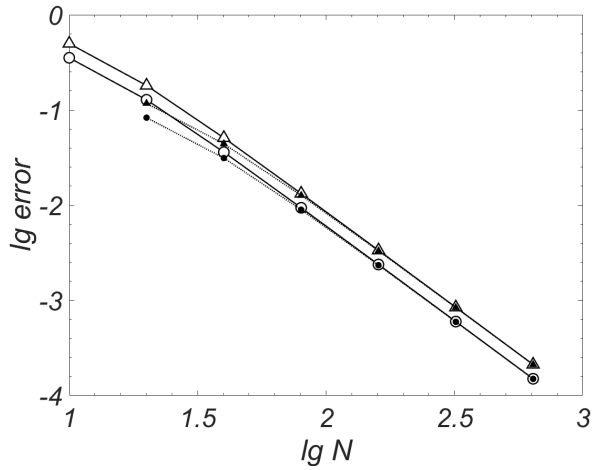

Рис. 7.10. Погрешность решения в задаче о падении p-поляризованной волны на диэлектрическую границу раздела. Обозначения соответствуют рис. 7.6.

На рис. 7.10 представлены погрешности сеточного расчета, масштаб графика двойной логарифмический. Как и в предыдущем расчете, видно, что разностная схема сходится со 2-м порядком точности. При этом на участке теоретической сходимости апостериорные оценки точности по методу Ричардсона-Калиткина (см. п. 1.4) практически неотличимы от фактических погрешностей, вычисленных при помощи прямого сравнения численного решения с точным.

7.5.3. Полное внутреннее отражение

1° Рассмотрим распространение света из оптически более плотной среды в оптически менее плотную. Будем считать, что обе среды не являются генерирующими. Как известно, если параметры среды таковы, что $\mathrm{n}_1/\mathrm{n}_2 \sin\alpha > 1$, то происходит полное внутреннее отражение. Волна полностью отражается обратно в плотную среду; по границе раздела распространяется поверхностная волна, амплитуда которой экспоненциально убывает по мере удаления от границы раздела.

2° Точное решение этой задачи дается формулами (7.34) – (7.37) для s-поляризации и (7.44) – (7.47) для p-поляризации. При этом в формулах (7.46), (7.47) нужно заменить $k \to ik$.

Аналогично строится точное решение методом оптических путей: в соотношениях (7.43) и (7.50) нужно сделать замену $k \to ik$.

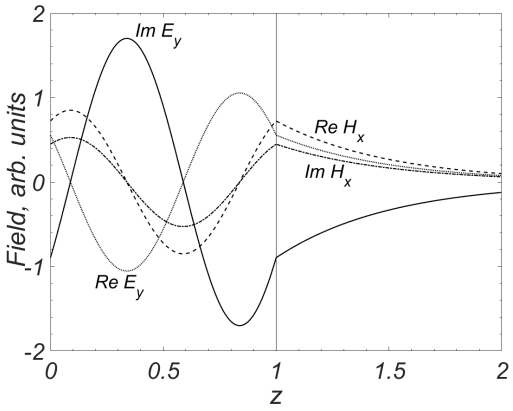

Рис. 7.11. Тангенциальные компоненты полей в задаче о полном внутреннем отражении s-поляризованной волны. Вертикальная прямая – граница раздела.

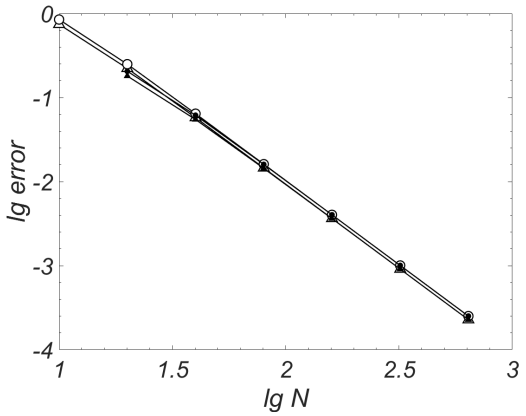

Рис. 7.12. Погрешность решения в задаче о полном внутреннем отражении s-поляризованной волны. Обозначения соответствуют рис. 7.6.

$\mathcal{З}^o$ Положим $a = 2$, $b = 1$, $\omega = 2\pi$, $\alpha = \pi/3$, $\varepsilon_1 = \mu_1 = 1$, $\varepsilon_2 = 0.1$, $\mu_2 = 1$, $c = 1$, $E^0 = 1$, $E^a = 0$. Приведем результаты численных расчетов для случая s-поляризации.

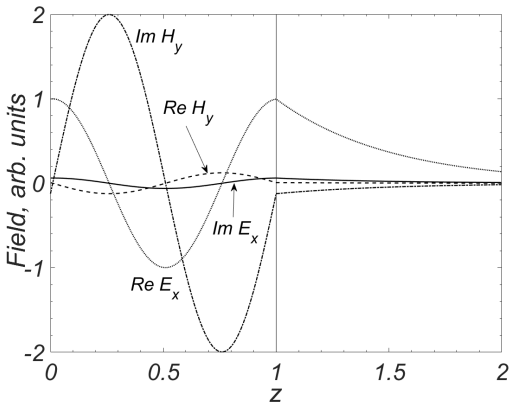

Рис. 7.13. Тангенциальные компоненты полей в задаче о полном внутреннем отражении p-поляризованной волны. Вертикальная прямая – граница раздела.

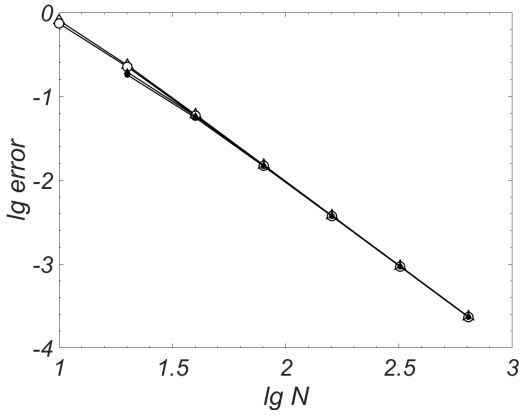

Рис. 7.14. Погрешность решения в задаче о полном внутреннем отражении p-поляризованной волны. Обозначения соответствуют рис. 7.6.

На рис. 7.11 представлены тангенциальные компоненты E_y и H_x. Видно, что они непрерывны, но имеют излом на границе раздела. Слева от границы раздела решение является осциллирующим, справа – имеет вид затухающей экспоненты. Такой вид решения является качественно разумным (напомним, что поля предполагаются неструктурированными).

На рис. 7.12 показаны погрешности, полученные в этом расчете, в зависимости от числа шагов сетки. Масштаб графика двойной логарифмический. Видно, что погрешности убывают в соответствии со 2-м порядком точности. На участке теоретической сходимости апостериорные оценки по методу Ричардсона-Калиткина (см. п. 1.4) практически неотличимы от истинных погрешностей, найденных как разность численного и точного решений.

4° На рис. 7.13 и 7.14 приведены результаты расчетов для p-поляризованной волны. Здесь справедливы те же выводы, что и в предыдущем расчете. Качественный вид решения (см. рис. 7.13) соответствует ожидаемому. Погрешность численного расчета (рис. 7.14) убывает в соответствии с теоретическим 2-м порядком точности и почти совпадает с апостериорными оценками по методу Ричардсона-Калиткина.

5° Расчеты, представленные в п. 7.5.1 – 7.5.3, верифицируют корректность реализации условий сопряжения (7.16), (7.17).

Замечание. Точные решения (7.34) – (7.37) и (7.44) – (7.47) соответствуют стационарной постановке. На их основе методом спектрального разложения легко построить решения соответствующих нестационарных задач, в которых падающая волна является не бесконечным цугом, а импульсом. Для этого в (7.34) – (7.37) и (7.44) – (7.47) в качестве E^0 нужно взять спектральную амплитуду падающего импульса и выполнить обратное преобразование Фурье.

Тогда точное решение нестационарной задачи будет иметь вид (7.34) – (7.37), (7.44) – (7.47), где вместо $E^0 \exp(iz\tilde{k}_{1,2}\cos\alpha)$, $E^0 \exp(-iz\tilde{k}_1\cos\alpha)$ нужно подставить соответственно $\mathcal{E}(\sqrt{\varepsilon_{1,2}\mu_{1,2}}z/c - t)$, $\mathcal{E}((2b-z)/c\sqrt{\varepsilon_1\mu_1} - t)$. Здесь $\mathcal{E}(t)$ – временна́я развертка падающего импульса. Найти эти решения в литературе

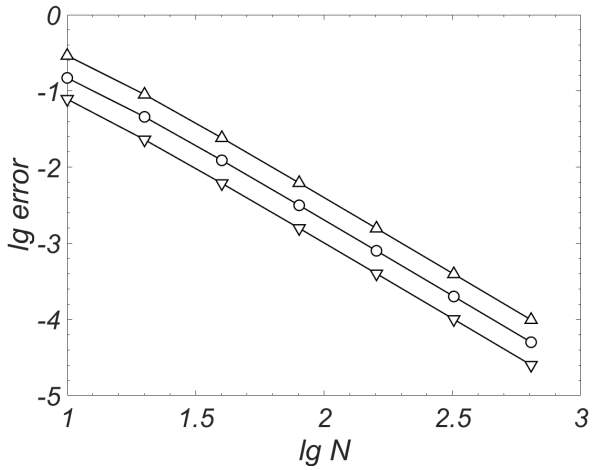

Рис. 7.15. Погрешность решения в задаче о законе Брюстера. ∇ – максимум модуля амплитуды отраженной волны. Остальные обозначения соответствуют рис. 7.6.

не удалось. По-видимому, они являются новыми. Расчеты этих задач не проводились, нестационарная формулировка метода оптических путей является простым обобщением стационарного варианта этого метода, а также метода спектрального разложения.

7.5.4. Эффект Брюстера

Пусть волна падает из воздуха $\varepsilon_1 = \mu_1 = 1$ на диэлектрик, у которого $\mu_2 = 1$. Если угол падения равен углу Брюстера, то отраженная волна полностью поляризована в плоскости, перпендикулярной плоскости падения. Поэтому в случае p-поляризации падающей волны отраженная волна будет отсутствовать.

В задаче из п. 7.5.2 положим $\mu_2 = 1$ и $\alpha = \operatorname{arctg}\sqrt{\varepsilon_2/\varepsilon_1}$. Слева от границы раздела решение представляет собой сумму падающей и отраженной волн. Чтобы найти отраженную волну E_{refl}, H_{refl}, вычтем из решения слева от границы раздела точное выражение для падающей волны, которое дается первыми слагаемыми в формулах (7.49).

Расчет проводился на сгущающихся сетках. На рис. 7.15 представлены погрешности полей E и H и максимум модуля амплитуды отраженной волны $\max|E_{\text{refl}}|$, $\max|H_{\text{refl}}|$ (в данном расчете эти максимумы для полей E и H оказались одинаковыми). Видно, что при амплитуда отраженной волны не пре-

восходит сеточной погрешности решения и убывает с той же скоростью при уменьшении шага сетки. Этот расчет показывает выполнение закона Брюстера для сеточного решения, что является дополнительной наглядной иллюстрацией разумности предложенного подхода (напомним, что поля предполагаются неструктурированными).

7.5.5. Интерферометр Фабри-Перо

1° Рассмотрим падение плоской волны на плоско-параллельную пластину (интерферометр Фабри-Перо). Пусть толщина пластины равна d, и границы раздела соответствуют плоскостям $z = b$ и $z = b + d$. Пластина расположена в воздухе. Материальные параметры пластины равны ε, μ.

2° Поскольку число переотражений внутри пластины формально бесконечно, то построение точного решения является затруднительным. Однако в этой задаче хорошо известен характер спектра отражения. Положим $\varepsilon = 4$, $\mu = 1$, $d = 0.5$, $c = 1$. Показатель преломления является вещественным, то есть пластинка оптически прозрачна. Тогда при нормальном падении нули в спектре отражения (то есть максимумы прохождения) соответствуют набегу фазы $\delta = 2\pi m$, $m = 1, 2, \ldots$, см. формулу (7.18). Соответствующие длины волн равны

$$\lambda = 2/m. \tag{7.51}$$

3° Процедура расчета спектра отражения описана в п. 6.1.5. На рис. 7.16 приведен расчетный спектр отражения от пластины при нормальном падении. Его качественный вид соответствует теоретическому. Контроль точности расчета проводился двумя способами.

Во-первых, вычислялась погрешность сеточного решения по методу Ричардсона-Калиткина. Погрешность на последней сетке, содержащей $N = 1280$ шагов, в зависимости от длины волны приведена на рис. 7.16. Видно, что с уменьшением длины волны (то есть с увеличением частоты) погрешность возрастает, поскольк-

ку увеличиваются производные решения. Тем не менее, даже для наименьшей из рассмотренных длин волн погрешность не превышает 1%.

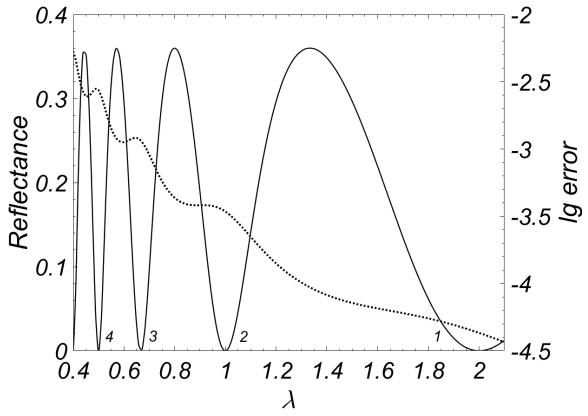

Рис. 7.16. Задача о прозрачной пластинке при нормальном падении. Сплошная линия – спектр отражения, пунктир – погрешность полей по методу Ричардсона-Калиткина на сетке с $N = 1280$. Цифры – номер минимума, равный значению m в (7.51).

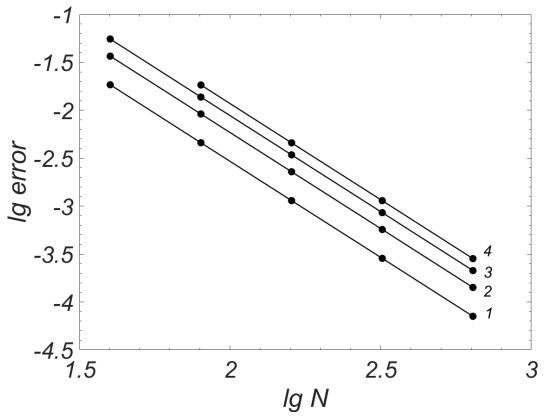

Рис. 7.17. Задача о прозрачной пластинке при нормальном падении. Погрешность положения m-го минимума в спектре отражения. Цифры около линий – значения m.

Во-вторых, вычислялась погрешность положений минимумов в спектре отражения. Она равна разности расчетных локальных минимумов и теоретических значений (7.51). Зависимость этой погрешности от числа шагов сетки для минимумов с $m = 1 \div 4$ приведена на рис. 7.17. Масштаб графика двойной логарифмический. Видно, с ростом m погрешности положения минимумов увеличиваются. При сгущении сеток все кривые выходят на прямые линии, то есть погрешности убывают по степенному закону. Скорость убывания соответствует 2-му порядку точности. Это означает, что, по крайней мере, в пределах рис. 7.17 погрешность положения минимума определяется только сеточной погрешностью разностной схемы. В противном случае погрешность вышла бы на некоторые предельные значения и дальше перестала бы убывать.

4° Аналогичные расчеты проводились при наклонном падении. Угол α подберем так, чтобы $\cos\beta$ было рациональным числом. Это позволит исключить

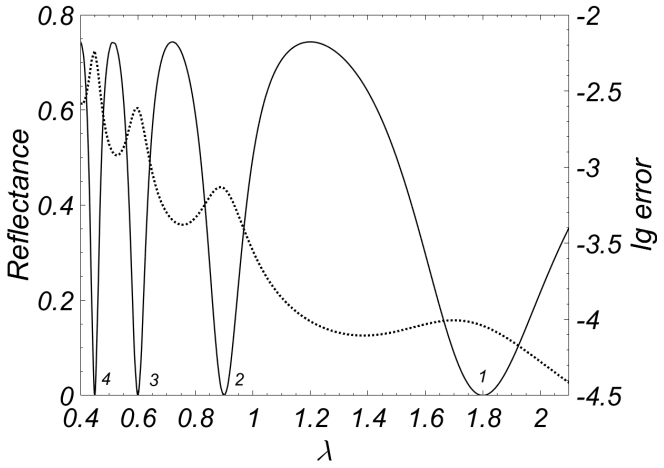

Рис. 7.18. Задача о прозрачной пластинке при наклонном падении. Обозначения соответствуют рис. 7.16.

влияние ошибок округления при вычислении тригонометрических функций и более аккуратно проверить сходимость. Положим $\cos\beta = 9/10$. Тогда $\alpha = \arcsin(\sqrt{\varepsilon}\cos\beta) \approx 1.059$. Пусть по-прежнему $\varepsilon = 4$, $\mu = 1$, $d = 0.5$, $c = 1$. Теоретические положения минимумов в спектре отражения есть $\lambda = 2\cos\beta/m = 9/(5m)$, $m = 1, 2, \ldots$

Полученный спектр отражения представлен на рис. 7.18. Его качественный вид полностью соответствует теоретически ожидаемому. Также на этом рисунке показаны апостериорные оценки погрешности полей на сетке с $N = 1280$ шагов, найденные по методу Ричардсона-Калиткина. Видно, что для всех рассмотренных длин волн погрешность решения составляет не более 1%. Зависимость погрешности положения минимумов от числа шагов сетки была почти неотличима от рис. 7.17.

Таким образом, если поглощение равно нулю (то есть пластинка прозрачна), то погрешность расчетного положения минимума определяется только погрешностью разностной схемы. При этом сам метод оптических путей не вносит физической погрешности.

5° Для материалов с поглощением метод оптических путей вносит некоторую физическую погрешность (см. п. 7.4.2). Она возрастает с увеличением $\mathrm{Im}\,\varepsilon$.

Чтобы оценить эту погрешность, проводились расчеты задачи о наклонном падении на пластинку с комплекснозначным ε.

В предыдущей задаче положим $\varepsilon = 4 + i$. Такое поглощение очень велико: типичные $\mathrm{Im}\,\varepsilon$ для актуальных материалов составляют $\sim 10^{-2} \div 10^{-3}$, реже 10^{-1} (см., например, рис. 6.13). Поэтому данный тест является представительным.

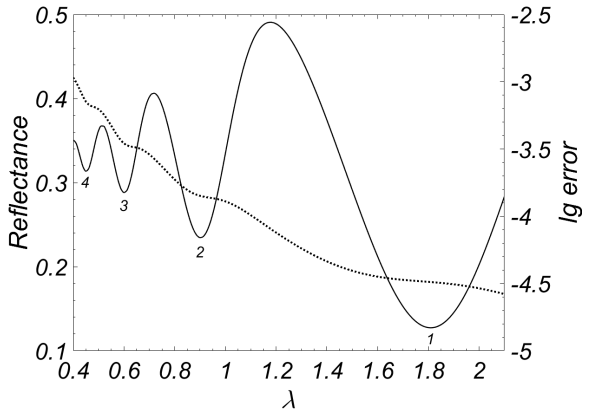

Рис. 7.19. Задача о поглощающей пластинке при наклонном падении. Обозначения соответствуют рис. 7.16.

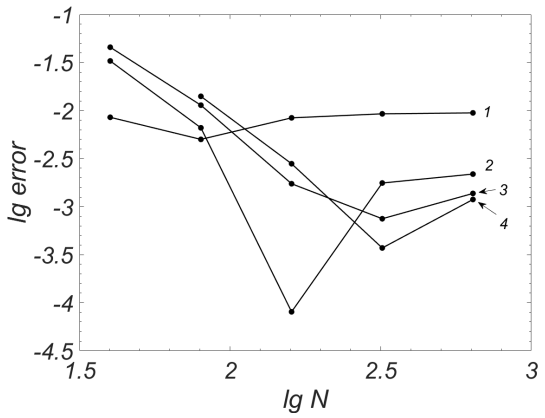

Рис. 7.20. Задача о поглощающей пластинке при наклонном падении. Погрешность положения m-го минимума в спектре отражения. Цифры около линий – значения m.

Расчетный спектр отражения приведен на рис. 7.19. Видно, что с увеличением m минимумы отражения становятся более мелкими. На этом рисунке показана также апостериорная оценка точности полей E и H на сетке с $N = 1280$ шагов. Поскольку решение является более плавным, чем в предыдущей задаче, то фактическая погрешность оказалась заметно меньше и не превосходит 0.3%.

На рис. 7.20 показана зависимость погрешности положения минимумов от числа шагов сетки. Масштаб графика двойной логарифмический. Вид кривых принципиально отличается от случая прозрачной среды. Начало кривых соответствует грубым сеткам. Здесь физическая погрешность метода оптических путей мала по сравнению с математической сеточной погрешность. Поэтому в начале кривых на рис. 7.20 погрешность убывает при увеличении числа шагов. Однако на достаточно подробных сетках сеточная погрешность становится

сопоставима с физической погрешностью. Поэтому при дальнейшем сгущении сетки погрешность на рис. 7.20 перестает убывать и выходит на константу. Эта константа различна для разных m. Она оказалась наибольшей для первого минимума $m = 1$. Однако даже для него физическая погрешность не превышает 1%. Такую точность можно считать отличной.

6° Таким образом, расчеты данного пункта верифицируют корректность приближения эффективных толщин и показывают, что метод оптических путей применим к широкому кругу важных прикладных задач.

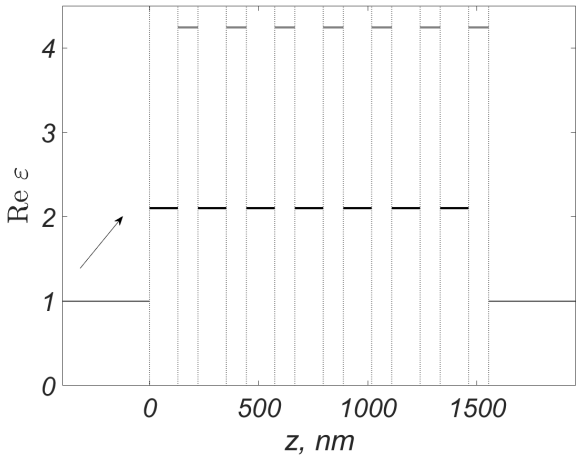

Рис. 7.21. Фотонный кристалл из работы [273]. Зависимость ε от координаты z, соответствующая длины волны на длине волны $\lambda = 900$ нм. Вертикальные линии – границы слоев. Стрелка – направление распространения падающей волны, угол падения равен 45^o.

7.5.6. Спектры фотонных кристаллов

Постановка задачи. Рассмотрим ФК, состоящий из 7 пар слоев $\{SiO_2 - 130$ нм, $Ta_2O_5 - 92$ нм$\}$ [273]. Эта структура приведена на рис. 7.21. Из $z = -\infty$ на структуру падает плоская линейно поляризованная монохроматическая волна. Угол падения равен 45^o.

Зависимость диэлектрических проницаемостей указанных материалов от длины волны показана на рис. 6.13, 6.14. В [273] опубликованы экспериментальные спектры отражения этого ФК при угле падения $\alpha = 45^o$ для s- и p-поляризаций.

Расчетные спектры. 1° Были вычислены спектры отражения и прохождения для волн обеих поляризаций. Сеточная погрешность расчета не превышала 0.1%.

Для s-поляризованной волны спектр отражения приведен на рис. 7.22, спектр прохождения – на рис. 7.23. Хорошо видна запрещенная зона в диапазоне длин волн 600÷850 нм; в этом диапазоне для данного угла падения ФК является зеркалом. Слева и справа от запрещенной зоны видны минимумы отражения, которые соответствуют практически полному прохождению.

Для p-поляризованной волны спектры отражения и прохождения приведены на рис. 7.24 и 7.25 соответственно. Видна запрещенная зона, соответствующая длинам волн $600 \div 800$ нм, и практически нулевые минимумы отражения слева и справа от нее.

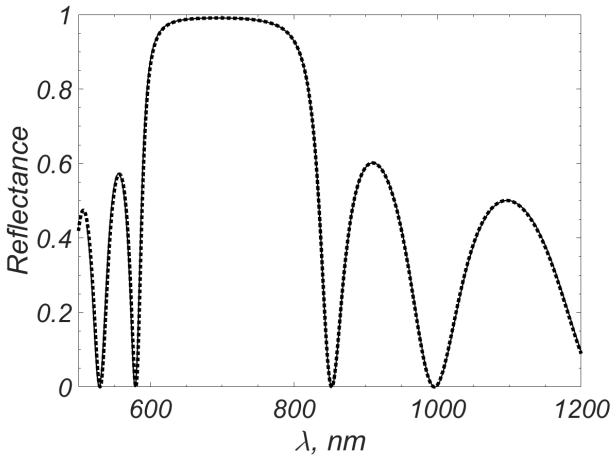

Рис. 7.22. Спектр отражения ФК на рис. 7.21, s-поляризованная волна. Сплошная линия – бикомпактная схема, пунктир – матричный метод.

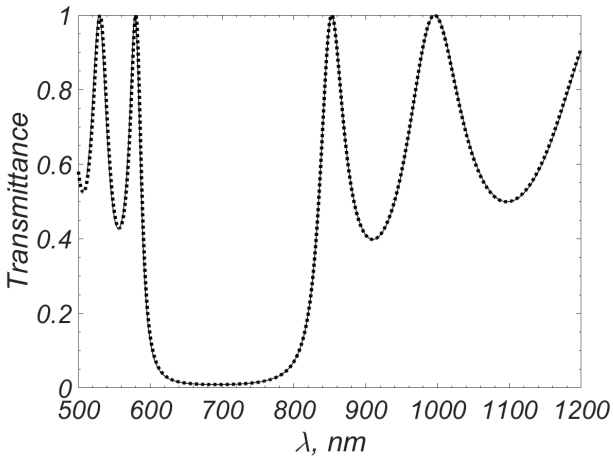

Рис. 7.23. Спектр прохождения ФК на рис. 7.21, s-поляризованная волна. Обозначения соответствуют рис. 7.22.

2° Для сравнения такие же расчеты были выполнен с помощью матричного метода [313,314]. Напомним, что соответствующие спектры являются точными. Они также приведен на рис. 7.22 – 7.25. Видно, что спектры, найденные разными методами, практически совпадают. Расхождение спектров, найденных с помощью бикомпатной схемы, и спектров, вычисленных с помощью матричного

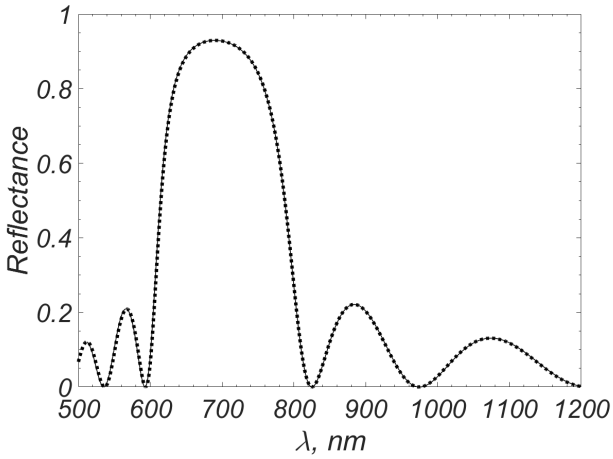

Рис. 7.24. Спектр отражения ФК на рис. 7.21, p-поляризованная волна. Обозначения соответствуют рис. 7.22.

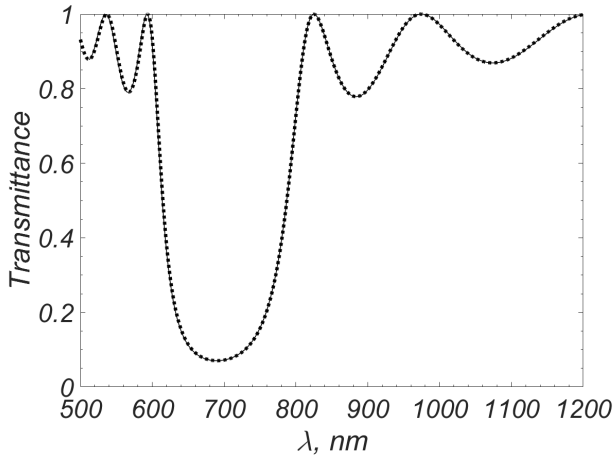

Рис. 7.25. Спектр прохождения ФК на рис. 7.21, p-поляризованная волна. Обозначения соответствуют рис. 7.22.

метода, не превышает 0.1%, что соответствует точности расчета по бикомпактной схеме. Это убедительно подтверждает правильность метода оптических путей.

Экспериментальные спектры. 1° Как отмечалось в п. 6.1.5, спектр ФК чувствителен к толщинам слоев. При этом из-за технологических ограничений фактические флуктуации толщин нельзя считать пренебрежимо малыми. Поэтому расчетный спектр идеального ФК (без учета флуктуаций) принципиально отличается от реального экспериментального спектра: вместо нулей отражения возникают ненулевые минимумы увеличенной ширины. При наклонном падении этот фактор проявляется сильнее, чем при нормальном.

Наличие систематической погрешности h_0 в толщинах (которую для простоты примем одинаковой для всех слоев) приводит к смещению минимумов в спектре отражения. Случайные погрешности (величина которых, вообще говоря, различна у разных слоев) приводит к размыванию минимумов и подъему их над осью абсцисс. Стандартное уклонение случайной погрешности обозначим через σ_0.

2° Чтобы учесть флуктуации толщин, применим метод виртуального эксперимента. Величина h_0 подбиралась по положению минимумов отражения, стан-

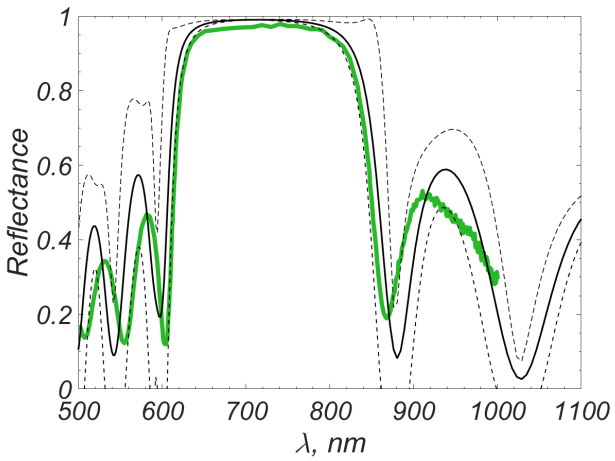

Рис. 7.26. Спектр отражения ФК на рис. 7.21, *s*-поляризованная волна. Жирная линия – эксперимент [273]. Тонкая линия – данная работа. Пунктир – границы доверительного интервала по методу виртуального эксперимента (соответствуют двум стандартным уклонениям).

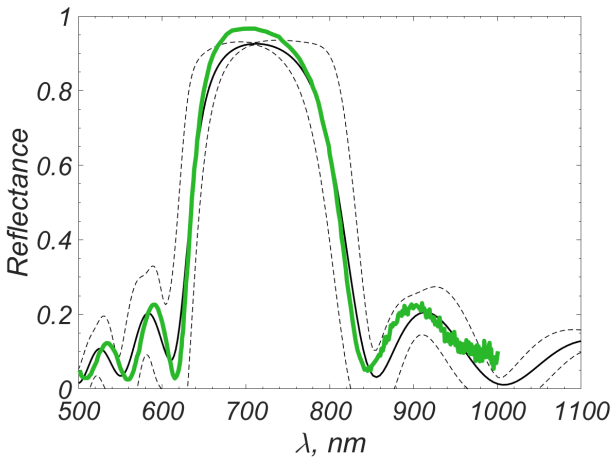

Рис. 7.27. Спектр отражения ФК на рис. 7.21, *p*-поляризованная волна. Обозначения соответствуют рис. 7.26.

дарт σ_0 – по их глубине. Для обеих поляризаций практически оптимальными оказались значения $h_0 \approx 3$ нм, $\sigma_0 \approx 3.5$ нм.

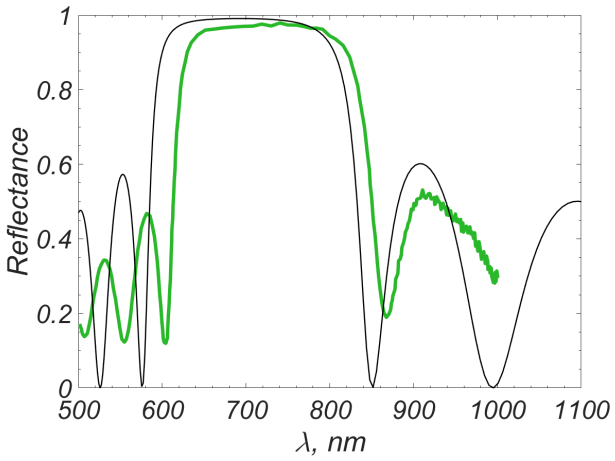

Рис. 7.28. Спектр прохождения ФК на рис. 7.21, *s*-поляризованная волна. Жирная линия – эксперимент [273]. Тонкая линия – расчет без усреднения по методу виртуального эксперимента.

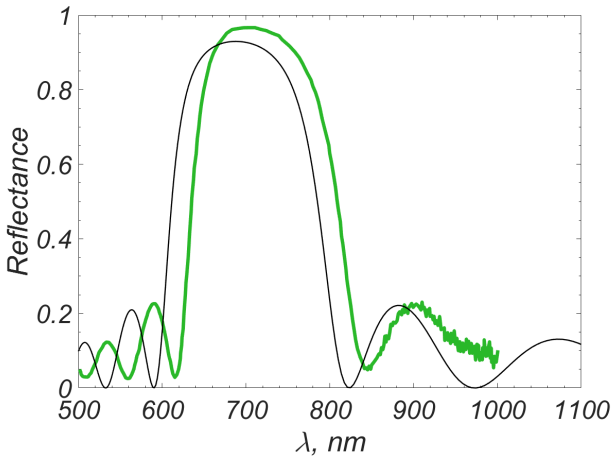

Рис. 7.29. Спектр прохождения ФК на рис. 7.21, *p*-поляризованная волна. Обозначения соответствуют рис. 7.28.

3° Сравнение полученного усредненного спектра с экспериментальным приведено на рис. 7.26 (для s-поляризации) и рис. 7.27 (для p-поляризации). Дополнительно показаны доверительные интервалы, равные двум стандартным уклонениям.

Видно, что положения экстремумов в расчетном и экспериментальном спектрах совпадают с точностью 1-2%. Такую точность можно считать отличной. Сами значения отражения передаются с точностью в основном 1-4%, изредка погрешность увеличивается до $\sim 10\%$. Это соответствует характерной точности экспериментальных измерений. Практически вся экспериментальная кривая лежит внутри доверительного коридора вокруг расчетной кривой. Это подтверждает разумность выбранных σ_0 и h_0.

4° Для сравнения на рис. 7.28, 7.29 приведены расчетные спектры отражения, полученные без использования метода виртуального эксперимента (то есть без учета флуктуаций толщин слоев). Видно, что имеет место качественное согласование, но количественная погрешность (расстояние между кривыми) достаточно велика. Это показывает, что применение метода виртуального эксперимента принципиально необходимо.

5° Проведенные расчеты хорошо согласуются с результатами эксперимента. Это является убедительным подтверждением правильности предложенных методов.

7.6. Поверхностные волны Блоха

7.6.1. Постановка задачи

1° Важное место занимают расчеты возбуждения поверхностных волн различного типа в одномерном фотонном кристалле. Эта задача имеет большое значение для экспериментальных исследований и перспективных технических приложений устройств интегральной фотоники, поскольку поверхностная волна может рассматриваться как носитель сигнала в оптических вычислительных

устройствах. От нее требуется максимально возможное время жизни. При этом, чтобы волну можно было детектировать в эксперименте, ее интенсивность свечения должна быть достаточно большой.

2° Пусть ФК из п. 6.1.5 расположен на границе раздела между стеклянной призмой и воздухом. Показатель преломления материала призмы равен 1.51. Между ФК и призмой находится слой иммерсионного масла с показателем преломления 1.49. Таким образом в слое $z < 0$ нужно положить $\varepsilon = 2.22$. Эта структура приведена на рис. 7.30.

Из призмы на ФК наклонно падает лазерный импульс лазера Ti:Sapphire Coherent MICRA5 [333]. Волна является плоской линейно поляризованной и имеет s-поляризацию. Спектр импульса приведен на рис. 7.31. Центральная длина волны равна 800 нм, ширина – 30 нм. Мы приближали спектр гауссовским профилем, он также приведен на рис. 7.31.

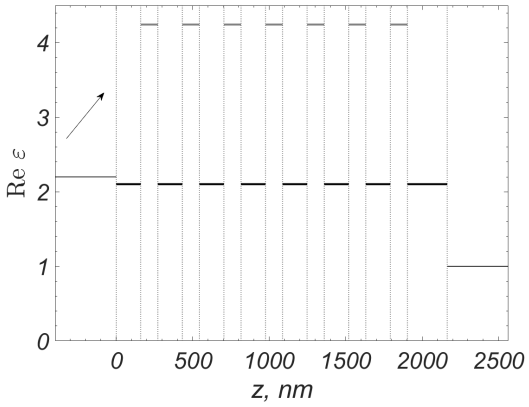

Рис. 7.30. К постановке задачи о динамике БПВ в фотонном кристалле. Зависимость ε от координаты z, соответствующая длины волны на длине волны $\lambda = 900$ нм. Вертикальные линии – границы слоев. Стрелка – направление распространения падающей волны, угол падения равен 45^o.

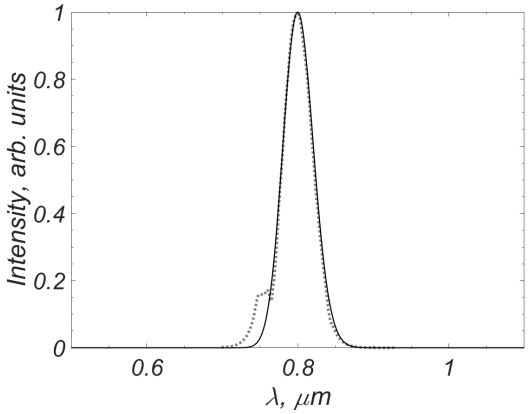

Рис. 7.31. Спектр падающего импульса. Пунктир – лазер MICRA5 [333], сплошная линия – аппроксимация гауссовым профилем.

3° На границе раздела раздела между последним слоем ФК и воздухом возникает полное внутреннее отражение. В фотонном кристалле формируется связанное состояние, которое называется блоховской поверхностной волной (БПВ). Энергия этого связанного состояния локализована внутри ФК (бóльшая часть – в последнем слое SiO_2). В водухе поля экспоненциально убывают по мере удаления от границы раздела «SiO_2-воздух».

В спектре отражения этому БПВ соответствует узкий провал (резонанс), форму которого можно считать лоренцевой. Эта волна распространяется вдоль поверхности фотонного кристалла, теряя энергию за счет излучения. Этот процесс принципиально нестационарный. Поэтому время жизни БПВ оказывается конечным. Напомним, что временем жизни такого связанного состояния называется время, за которое интенсивность свечения уменьшается в e раз.

Задачей данного раздела является расчет времени жизни связанного состояния и исследование зависимости времени жизни от толщин слоев фотонного кристалла.

7.6.2. Выбор сетки по частоте

1° Серьезной проблемой при расчете является наличие узких резонансов в спектре отражения. Например, спектральная ширина импульса лазера MICRA составляет ≈ 30 нм. Чтобы адекватно передать импульс, нужно брать ширину частотного диапазона в ~ 5 раз больше ширины импульса, то есть ~ 150 нм. При этом характерная ширина резонансов ФК при больших углах падения составляет от 0.5 нм до 0.02 нм, то есть равномерная сетка по частоте должна содержать $10^4 - 10^5$ шагов. Для каждой из этих частот нужно решать стационарную задачу, сводящуюся к решению системы линейных алгебраических уравнений достаточно большой размерности. Такая трудоемкость неприемлема.

Если же сетка по частоте недостаточно подробная, и область резонанса не разрешена, то характер сеточного решения качественно отличается от точного. Сеточные гармоники, представленные в падающем импульсе, не резонируют

со структурой, и в связанное состояние попадает существенно меньшая энергия (либо связанное состояние вовсе не образуется). Это напоминает известный стробоскопический эффект.

2° Чтобы преодолеть эту трудность, была разработана процедура автоматического выбора шага по частоте, ориентированная на расчеты задач с полным внутренним отражением.

Вместо исходной нестационарной задачи решается набор стационарных задач (относительно спектральных амплитуд). Зададим диапазон частот, учитываемых в падающем импульсе (так, чтобы за пределами этого диапазона спектральные амплитуды были заведомо пренебрежимо малы). Решим стационарную задачу для наименьшей из частот. Одновременно с решением найдем соответствующий коэффициент отражения по интенсивности R. Следующую частоту выберем по формуле

$$\omega_{n+1} = \omega_n + \frac{h_\omega}{1 + A(1 - R_n)}. \tag{7.52}$$

Здесь A и h_ω – настроечные параметры. Если $R = 1$, то шаг по частоте равен h_ω. Если $R \neq 1$, то шаг уменьшается тем сильнее, чем глубже «провал» в спектре отражения. Параметр A целесообразно выбирать по добротности резонанса.

Выполним расчет стационарной задачи для ω_{n+1} и найдем для нее R_{n+1}. Для выбора следующей частоты ω_{n+2} снова применим формулу (7.52) и т.д. Такие расчеты проводятся до тех пор, пока очередное ω_n не станет больше правой границы выбранного спектрального диапазона.

3° Выбор шага по частоте по описанному алгоритму позволяет а) надежно разрешать стенки и дно резонансных минимумов (даже если они очень узкие); б) повышать точность расчета; в) обнаруживать малозаметные минимумы в спектре отражения по существенному изменению шага по частоте.

7.6.3. Расчет времени жизни БПВ

1° Время жизни связанного состояния определяют по кросс-корреляционной функции. Последняя равна свертке интенсивностей отраженного I_{refl} и референсного (падающего) I_{inc} импульсов.

$$C(t) = \int_{-\infty}^{+\infty} I_{\text{inc}}(\xi) I_{\text{refl}}(t-\xi) d\xi. \tag{7.53}$$

2° Падающий импульс имеет гауссов профиль. Если поверхностная волна не формируется, то отраженный импульс является гауссовым. Тогда $C(t)$ также имеет гауссов профиль.

3° Если поверхностная волна сформировалась, то профиль отраженного импульса отличается от падающего. Характерный вид временной развертки интенсивностей падающего и отраженного импульсов приведен на рис. 7.32. Отраженный импульс является асимметричным, причем правое крыло затухает экспоненциально (то есть намного медленнее, чем левое крыло). Это экспоненциальное затухание соответствует излучению поверхностной волны. В этом случае $C(t)$ имеет гауссову «макушку» и экспоненциально затухающий «хвост» $I \sim \exp(-t/\tau)$, где τ – время жизни поверхностной волны. В полулогарифмическом масштабе этот «хвост» превращается в прямую линию, наклон которой равен $1/\tau$. Отсюда нетрудно найти τ.

4° Из-за сеточной погрешности «хвост» кросс-корреляционной функции в полулогарифмическом масштабе оказывается не строго прямолинейным. Это может внести заметную погрешность при определении времени жизни (особенно, если наклон этого прямолинейного участка мал). В данной работе построена следующая процедура.

Проведем расчет кросс-корреляционной функции $C(t)$ с одновременной оценкой погрешности $\delta C(t)$ по методу Ричардсона-Калиткина. Результатом расчета является массив сеточных значений C_n этой функции и массив поточечных оценок погрешности δC_n. Выберем визуально прямолинейный участок «хвоста» и методом наименьших квадратов аппроксимируем его линейной функцией с ве-

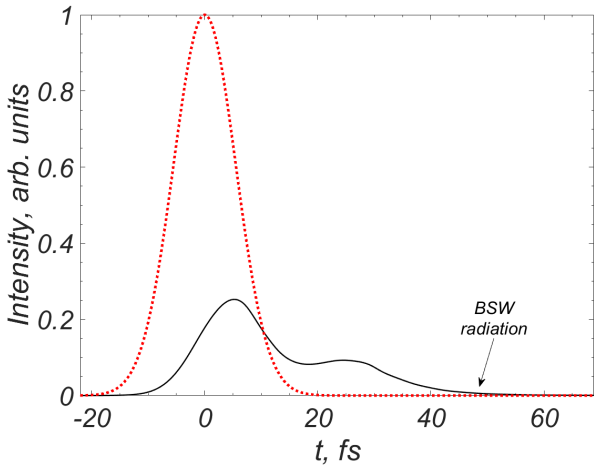

Рис. 7.32. Характерная временная развертка интенсивности падающего импульса (пунктир) и отраженного импульса (сплошная линия).

сами, обратно пропорциональными квадратам поточечных погрешностей

$$C_n \approx At_n + B, \quad \sum_n \delta C_n^{-2}(At_n + B - C_n)^2 \to \min_{A,B}. \tag{7.54}$$

По наклону этой прямой вычислим время жизни поверхностной волны. Проводя дальнейшие сгущения сетки, вычислим оценку точности для времени жизни по методу Ричардсона-Калиткина и применим экстраполяцию погрешности. При этом контролируется сходимость как самих расчетных значений τ, так и результатов экстраполяции. Описанная процедура значительно повышает точность и достоверность вычисления времен жизни стоячей волны. В задачах фотоники такой подход ранее не использовался.

7.6.4. Спектр отражения

1° На рис. 7.33 приведены нормированные спектры отражения, равные спектру отражения, деленному на спектр падающего импульса. Расчеты проводились для углов падения $\alpha_0 = 43^o$, 44^o, 45^o, 46^o. Точность расчета спектров составила не менее 0.01%. В расчетах не учитывались флуктуации толщин слоев, поскольку они зависят от качества изготовления конкретного образца.

Видно, что все спектры отражения имеют узкий резонанс, соответствующий формированию связанного состояния. Резонансную длину волны обозна-

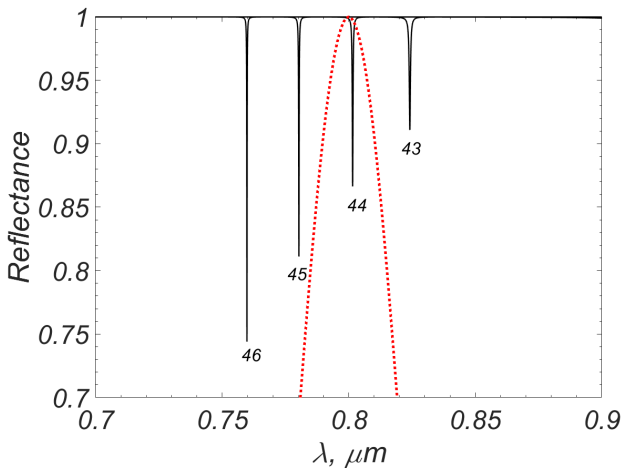

Рис. 7.33. Спектры отражения от ФК (сплошные линии). Цифры около кривых – углы падения в градусах. Пунктир – спектр падающего импульса.

чим через λ_0. Увеличение α приводит к уменьшению λ_0. При этом добротность резонанса увеличивается: глубина увеличивается, а ширина уменьшается. Поэтому можно ожидать, что при $\alpha = 43^o$ время жизни будет наименьшим, а при $\alpha = 46^o$ – наибольшим.

2° Для сравнения на рис. 7.33 приведен спектр падающего импульса. Видно, что резонанс, соответствующий углу $\alpha = 44^o$, расположен практически у максимума λ_0 этого спектра. Это означает, что при данном α в связанное состояние передается наибольшая энергия падающего поля, и можно ожидать наиболее интенсивного излучения поверхностной волны. Резонанс, соответствующий углу $\alpha = 43^o$, расположен правее λ_0. Резонансы, соответствующие $\alpha = 45^o$ и 46^o, расположены левее λ_0. Полученные положения и глубины резонансов приведены в табл. 7.1.

3° Чтобы проиллюстрировать работу алгоритма выбора шага по частоте, приведем участок спектра для $\alpha = 46^o$ вблизи резонанса (см. рис. 7.34). Маркерами отмечены узлы по длине волны. Видно, что по мере приближения к минимуму отражения шаг сгущается. Вдали от минимума шаг довольно быстро увеличивается. Видно также, что на склоны и дно резонанса попадает много точек, и ширина и глубина определяются надежно (стробоскопический эффект отсутствует). Аналогичная картина имела место и при других углах падения.

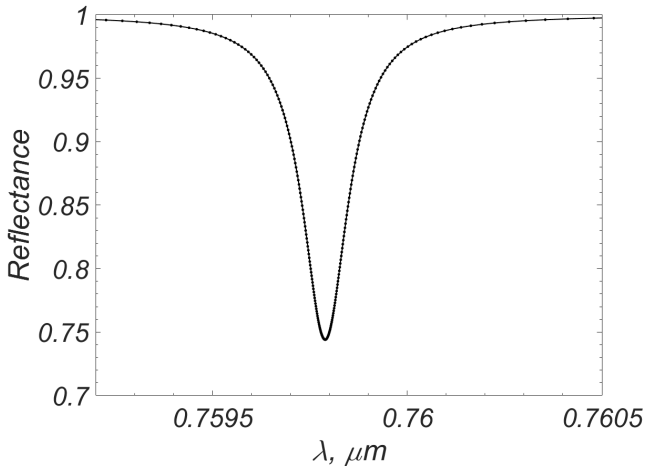

Рис. 7.34. Спектр отражения от ФК при $\alpha = 46^o$. Маркеры – узлы сетки по длине волны.

Таблица 7.1. Динамика БПВ в фотонном кристалле.

α, ^{o}	λ_0, мкм	$\min R$	$\lg C_0$, отн. ед.	$\delta \lg C$, отн. ед.	τ, фс	$\delta\tau$, фс
43	0.8241	0.9108	-2.9	-4.1	665.8	0.6
44	0.8015	0.8664	-2.8	-4.3	1030	4
45	0.7803	0.8111	-3.4	-4.7	1545	8
46	0.7598	0.7440	-4.8	-4.5	2190	11

4° Сравним результаты расчетов с экспериментально измеренными спектрами отраженного импульса. Их любезно предоставили автору Бессонов и Попкова, входящие в состав коллектива под руководством Федянина на кафедре квантовой электроники физического факультета МГУ им. М.В. Ломоносова. Эти спектры приведены на рис. 7.35, 7.36.

Экспериментальные измерения спектров отражения проводились для двух образцов ФК. Для каждого из них доступно несколько измерений, в которых луч направлялся на разные точки образца. Мы усреднили эти данные и вычислили среднеквадратичное уклонение. Последнее принимается за оценку погрешности эксперимента.

От измерения к измерению спектр падающего импульса несколько отличался. Во-первых, отличались центральные длины волн. Во-вторых, при под-

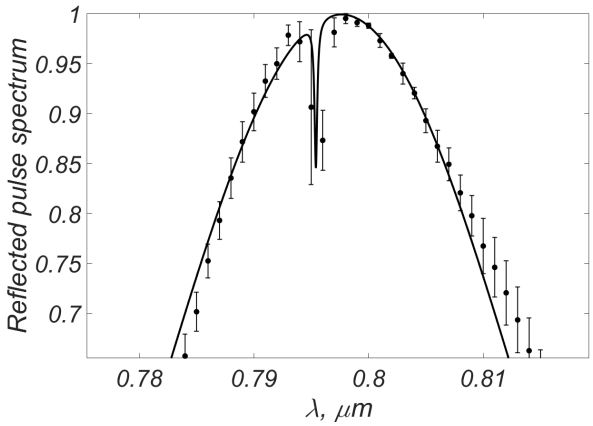

Рис. 7.35. Ненормированный спектр отражения в образце 1. Точки – эксперимент, кривая – данная работа.

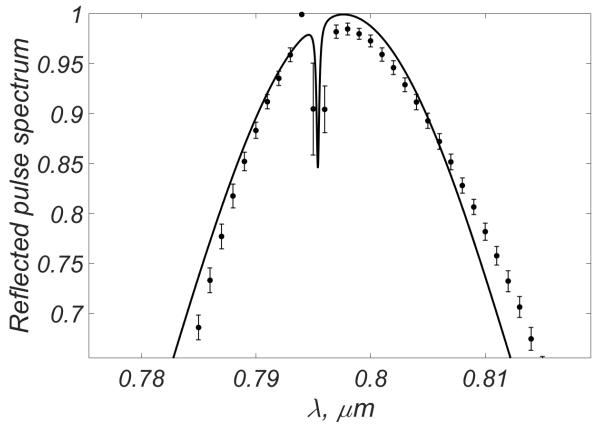

Рис. 7.36. Ненормированный спектр отражения в образце 2. Обозначения соответствуют рис. 7.35.

стройке центральной длины волны несколько деформируются «склоны» спектра. Чтобы учесть этот фактор, мы варьировали центральную длину волны Λ_0, а ширина спектра на половине высоты $\delta\Lambda = 32$ нм была одинакова во всех расчетах.

Также в разных измерениях несколько отличался угол падения луча на ФК. Поэтому в расчете угол падения подбирался для наилучшего описания положения резонанса.

Расчетные спектры также приведены на рис. 7.35, 7.36 Видно, что они согласуются с экспериментальными данными в пределах указанной погрешности. Это подтверждает адекватность используемых табличных данных для диэлектрических проницаемостей SiO_2 и Ta_2O_5.

7.6.5. Кросс-корреляционная функция

1º На рис. 7.37 приведена зависимость кросс-корреляционных функций от временного сдвига отраженного импульса относительно падающего (см. формулу (7.53)) для всех углов падения. Для наглядности график построен в полулогарифмическом масштабе. Излучению поверхностной волны соответствует

прямолинейный участок графика. Жирной линией отмечен участок, по которому вычислялось время жизни связанного состояния.

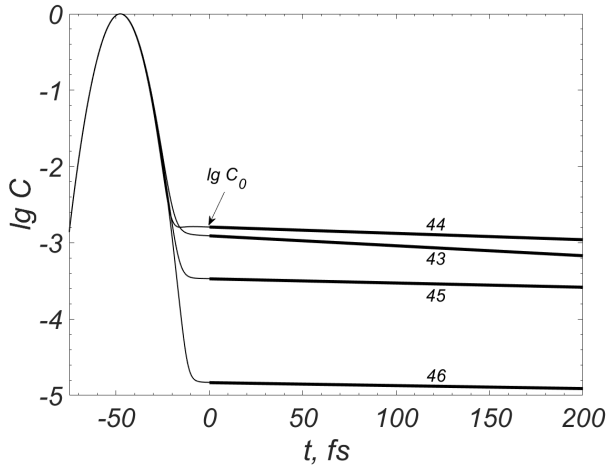

Рис. 7.37. Десятичный логарифм кросс-корреляционной функции. Цифры около линий – углы падения. Жирным выделены прямолинейные участки, по которым вычислялось время жизни БПВ.

2° В эксперименте (см., например, [273]) непосредственно измеряется кросс-корреляционная функция. Чем больше ее величина на прямолинейном участке рис. 7.37, тем больше интенсивность излучения БПВ. Поэтому за количественную характеристику яркости БПВ примем величину $\lg C_0$ в начале прямолинейного участка. Значения $\lg C_0$ для всех 4 углов, а также средние погрешности вычисления кросс-корреляционной функции $\delta \lg C$ приведены в табл. 7.1.

Чтобы детектировать БПВ в эксперименте, интенсивность ее излучения должна быть достаточно велика. На практике значение $C(t)$ на участке, соответствующем БПВ, должно составлять не менее $0.03 \div 0.1\%$ от максимального значения $C(t)$ (то есть $\lg C_0 \geqslant -3.5 \div -3$). Из табл. 7.1 видно, что БПВ является детектируемой при $\alpha = 43^o \div 45^o$. При этих углах средние погрешности существенно меньше самих значений $C(t)$, и расчет «хвоста» кросс-корреляционной функции является надежным. При $\alpha = 46^o$ средняя погрешность расчета $C(t)$ сопоставима со значениями $C(t)$ на «хвосте» на рис. 7.37, и проведенный расчет можно считать только оценочным. Однако интенсивность излучения БПВ настолько

мала, что наблюдать БПВ в эксперименте не удастся. Поэтому более точный расчет не проводился.

3° Найденные времена жизни БПВ приведены в табл. 7.1. Видно, что с увеличением угла время жизни увеличивается и достигает ~ 1500 фс для $\alpha = 45^o$. Это значение в ~ 50 раз превосходит времена жизни таммовских состояний в плазмонных структурах [273]. Поэтому полностью диэлектрические фотонные кристаллы, в которых реализуются БПВ, исключительно перспективны для технических приложений. Погрешности найденных времен жизни также приведены в табл. 7.1. Они составляют от $0.1 \div 0.5\%$. Такая точность существенно превосходит мировой уровень и заведомо перекрывает потребности практики.

7.6.6. Зависимость динамики БПВ от толщин слоев ФК

1° Для приложений актуален поиск таких толщин слоев ФК, при которых время жизни БПВ является наибольшим при хорошей детектируемости в эксперименте. Для этого был проведен расчет спектра отражения, кросс-корреляционной функции и времени жизни БПВ для различных толщин слоев ФК. Угол падения равнялся $\alpha = 45^o$.

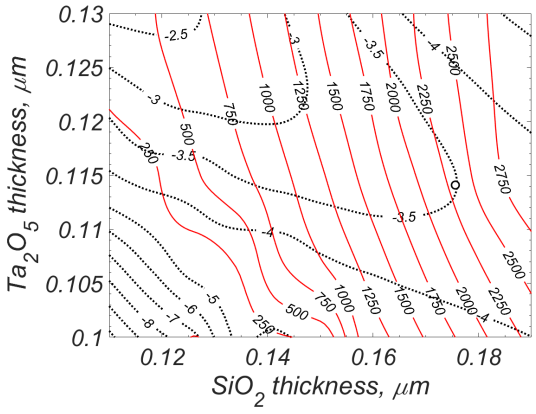

Рис. 7.38. Величина $\lg C_0$, отн. ед. (пунктир) и время жизни связанного состояния τ, фс (сплошные линии) в зависимости от толщин слоев ФК. $\circ$ – оптимальные толщины слоев.

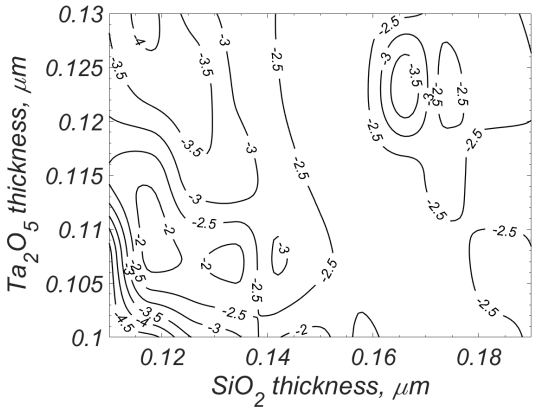

Рис. 7.39. Десятичные логарифмы относительных погрешностей вычисления времени жизни БПВ в зависимости от толщин слоев ФК.

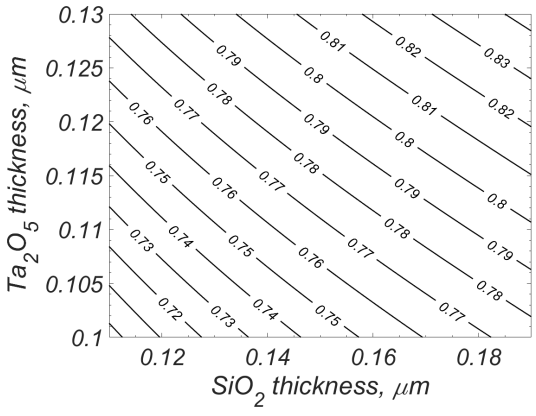

Рис. 7.40. Зависимость положения минимума в спектре отражения λ_0, мкм от толщин слоев ФК.

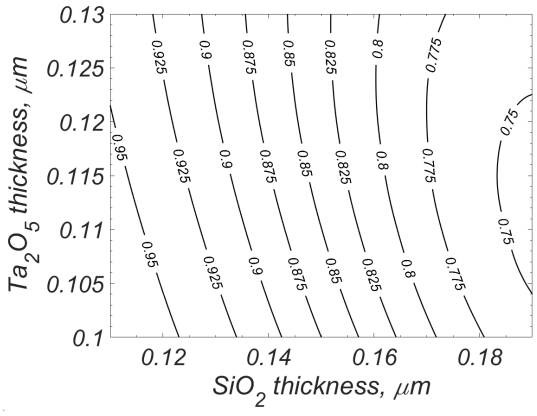

Рис. 7.41. Зависимость глубины минимума в спектре отражения $\min R$ от толщин слоев ФК.

2° На рис. 7.38 приведена зависимость $\lg C_0$ и τ от толщин слоев фотонного кристалла $h_{\mathrm{SiO_2}}$, $h_{\mathrm{Ta_2O_5}}$. С учетом указанных выше экспериментальных ограничений задача сводится к отысканию условного максимума $\max\tau$ на множестве $\{h_{\mathrm{SiO_2}}, h_{\mathrm{Ta_2O_5}} : \lg C_0 \geqslant -3.5\}$. Эта задача оптимизации сложна, поскольку множество ограничений задается не аналитически, а само является результатом численного решения. Мы решали эту задачу графически с помощью рис. 7.38. Строгое решение этой задачи выходит за рамки данной работы.

3° Из рис. 7.38 видно, что τ достаточно сильно зависит от $h_{\mathrm{SiO_2}}$ и сравнительно слабо – от $h_{\mathrm{Ta_2O_5}}$, то есть изолинии τ близки к вертикальным. Величина $\lg C_0$ имеет вид хребта, высота которого понижается из правого верхнего угла графика к левому нижнему углу. Видно, что $\max\tau$ достигается на границе указанного множества. Он составляет ≈ 2320 фс. Соответствующие толщины слоев равны $h_{\mathrm{SiO_2}} = 176$ нм, $h_{\mathrm{Ta_2O_5}} = 114$ нм.

4° На рис. 7.39 приведены десятичные логарифмы относительных погрешностей расчета времени жизни $\lg(\delta\tau/\tau)$. Видно, что на рассматриваемом множестве ограничений эта погрешность составляет не более 1%. Такая точность существенно превосходит мировой уровень. Она заведомо достаточна для приложений.

5° На рис. 7.40 и 7.41 приведены соответственно положение λ_0 и глубина $\min R$ минимума в спектре отражения. Видно, что λ_0 возрастает из левого нижнего угла в правый верхний, причем эта зависимость практически линейная: $\lambda_0 \approx Ah_{\mathrm{SiO_2}} + Bh_{\mathrm{Ta_2O_5}}$, A, B – некоторые числа. Это естественно, поскольку разность фаз между лучом, отраженным от поверхности ФК, и лучом, прошедшим туда и обратно через всю структуру, линейно растет с увеличением толщин слоев.

Глубина минимума возрастает с увеличением $h_{\mathrm{SiO_2}}$, а ее зависимость от $h_{\mathrm{Ta_2O_5}}$ оказывается сравнительно слабой. Наибольшая глубина равна $\min R \approx 0.75$, наименьшая – $\min R \approx 0.95$.

6° Полученные результаты могут непосредственно использоваться при планировании новых экспериментов по реализации долгоживущих связанных состояний в полностью диэлектрическом фотонном кристалле. Поэтому они представляют самостоятельную практическую ценность.

7.7. Основные результаты главы

1. Для задачи о наклонном падении плоской волны на одномерный фотонный кристалл предложен новый метод интегрирования уравнений Максвелла вдоль оптического луча (метод оптических путей). Этот метод фактически сводит эту задачу к одномерной. Предложенный подход позволяет существенно снизить трудоемкость решения многих важных задач.
2. Проведены расчеты тестовых задач с известным точным решением (наклонное падение плоской волны на плоскую границу раздела и интерферометр Фабри-Перо). Эти расчеты убедительно верифицируют метод оптических путей.
3. Проведены расчеты спектров отражения и прохождения диэлектрических фотонных кристаллов. Выполнено сравнение расчетов по предложенным методам и по методу матриц рассеяния. Показано, что расхождение соот-

ветствует сеточной погрешности расчета по бикомпактной схеме и методу оптических путей. Эти расчеты также убедительно верифицируют предложенные методы.

4. Проведены расчеты реальной задачи о рассеянии света на полностью диэлектрическом фотонном кристалле, опубликованном в работе [273]. Проведено сравнение расчетных спектров с ранее опубликованными экспериментальными. Показано, что разработанные методы (бикомпактные схемы, метод оптических путей и метод виртуального эксперимента) обеспечивают хорошее согласование расчетов с экспериментом в пределах точности последнего $1 \div 7\%$.
5. Проведены расчеты реальной задачи о формировании поверхностной волны Блоха при наклонном падении импульса на одномерный фотонный кристалл. Проведено исследование параметров этой волны (яркость свечения и время жизни связанного состояния) в зависимости от толщин слоев фотонного кристалла. Результаты этих расчетов можно использовать для экспериментальной реализации долгоживущих связанных состояний. Методы, предложенные в данной работе, позволили существенно повысить точность этого исследования.
6. Предложенные методы обеспечивают высокую количественную точность расчета реальных нестационарных задач фотоники. Это особенно важно при численных исследованиях сложных эффектов, таких как долгоживущие связанные состояния (поверхностные волны различных типов), а также при разработке новых устройств интегральной фотоники.

8. Скорости реакций

Помимо вопросов, перечисленных в предыдущих главах, соискателем рассмотрены задачи обработки экспериментальных данных, измеренных со значительными погрешностями. Типичными примерами таких данных являются скорости химических и термоядерных реакций.

Приложения. При моделировании процессов в высоко- и низкотемпературной плазме важными входными данными являются модели скоростей реакций. Для расчета энерговыхода в мишенях управляемого термоядерного синтеза (УТС) и неустойчивостей в токамаках требуются данные по скоростям термоядерных реакций с участием изотопов H и He. Газодинамические расчеты с учетом газофазных химических реакций возникают в задачах пиролиза углеводородов, при моделировании вхождения спускаемых аппаратов в атмосферу, при расчетах эволюции в атмосфере водорода, испускаемого из земных глубин (что влияет на формирование погоды и озоновых дыр) и ряда других задач.

Математический метод. Соискателем предложены новые методы аппроксимации экспериментальных данных, измеренных с большими погрешностями. Эти методы основаны на регуляризации специфического переопределенного ряда Фурье и на построении системы функций, ортогональных на заданном наборе точек с произвольными весами. Разработаны математические процедуры, позволяющие надежно находить доверительные интервалы этих аппроксимаций. Ранее такие оценки были неизвестны.

Термоядерные реакции. Соискателем уточнен список реакций, существенных для проблемы УТС. Традиционно учитываются реакции D+D→p+T, D+D→n+^{3}He, D+T→n+^{4}He, D+^{3}He→n+^{4}He. Показано, что при низких температурах (то есть на начальном этапе зажигания мишеней) заметными являются реакции T+T→2n+^{4}He, D+p→ γ+^{3}He, T+p→ γ+^{4}He.

К указанным 7 реакциям применены новые методы аппроксимации. Получены таблицы скоростей этих реакций в заведомо достаточном диапазоне температур от 10 эВ до 2 МэВ. Достигнутая точность составляет 1-4% при малых

температурах и доли процента при больших температурах, что в ~ 5 раз точнее известных ранее формул. Эти данные надежно перекрывают потребности расчетов лазерных мишеней и токамаков. Они переданы 2 коллективам газодинамиков из Института прикладной математики им. М.В. Келдыша РАН для практического использования.

Химические реакции. Рассмотрены 20 реакций водородо-воздушного горения, существенных для задач плазмохимии (т.е. при температурах до 1-2 кК и давлениях 1 атм.) и 15 реакций, описывающих автокаталитический механизм пиролиза этана при давлениях от 1 до нескольких атмосфер. Собран компендиум прямых экспериментальных данных по скоростям указанных реакций. Ранее такие компендиумы не создавались. К указанным реакциям применены предложенные математические методы обработки экспериментальных данных. Получены таблицы коэффициентов обобщенной формулы Аррениуса с оценками доверительных интервалов. Точность нахождения скоростей реакций составляет от долей процента до $\sim$20%, что в 3-5 раз точнее известных ранее формул.

База данных. Полученные данные размещены в базе данных ТЕФИС, разрабатываемой в Институте прикладной математики им. М.В. Келдыша РАН.

Данные результаты не выносятся на защиту.

9. Заключение

В данной работе получены следующие результаты.

1. Разработан, обоснован и протестирован алгоритм выбора шага по кривизне интегральной кривой для численного интегрирования задач Коши для ОДУ. Построены экономичные одношаговые схемы для вычисления кривизны одновременно с решением ОДУ. Разработана процедура сгущения адаптивных сеток, позволяющая вычислять асимптотически точную оценку погрешности по методу Ричардсона. Предложенные методы реализованы в виде комплекса проблемно-ориентированных программ. Проведено сравнение программ Гира и Дормана-Принса с предложенными методами и показаны преимущества последних. Проведено численное моделирование прикладной задачи кинетики реакций водород-кислородного горения.
2. Разработана, программно реализована и протестирована явная схема для расчета кинетики реакций, которая имеет второй порядок точности и обеспечивает неотрицательность численного решения. Выполнено сравнение предложенной схемы с другими методами первого и второго порядка точности и показаны ее преимущества в задачах кинетики реакций.
3. Разработаны, обоснованы и протестированы методы численного исследования подвижных особых точек и их последовательностей в решении ОДУ с апостериорной асимптотически точной оценкой погрешности. Предложенные методы реализованы в виде комплекса проблемно-ориентированных программ.
4. Предложенные методы применены к исследованию сингулярностей в решениях уравнений в частных производных методом прямых.
5. Разработана, обоснована, программно реализована и протестирована бикомпактная разностная схема для одномерной стационарной системы уравнений Максвелла в слоистых средах. Проведено сравнение предложенных схем с другими методами (МКЭ, МКР во временной области) и показаны преимущества предложенных подходов.

6. Разработан, обоснован и протестирован метод интегрирования уравнений Максвелла вдоль оптического луча для задачи о наклонном падении плоской волны на набор плоско-параллельных пластин. Предложенный метод реализован в виде комплекса проблемно-ориентированных программ.

Диссертация основана на идеях школы члена-корреспондента РАН Н.Н. Калиткина. Он определял общее направление работ и первоначальные постановки ряда задач, решенных в диссертации, а также участвовал в обсуждении результатов. Научный консультант Л.А. Севастьянов участвовал в написании текста диссертации и обсуждении результатов.

Список иллюстраций

Список таблиц

Список литературы

[1] А.П. Юшкевич. История математики. Т. 2. — М.: Наука, 1970.

[2] А.Н. Крылов. Лекции о приближенных вычислениях. — Л.: Изд-во АН СССР, 1933.

[3] Дж.Н. Ватсон. Теория бесселевых функций. — М.: ИЛ, 1949.

[4] В.В. Голубев. Лекции по аналитической теории дифференциальных уравнений. — М.-Л., ГИТТЛ, 1941.

[5] Э. Хайрер, С. Нерсет, Г. Ваннер. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — М.: Мир, 1990.

[6] C. Runge, H. König. Vorlesungen über Numerisches Rechnen. — Springer-Verlag, 1924.

[7] Potashov I.M., Tchemarina Ju.V., Tsirulev A.N. Bound orbits near scalar field naked singularities // European Physical Journal C. — 2019. — Vol. 79. — P. 709–719.

[8] Potashov I.M., Tchemarina Ju.V., Tsirulev A.N. Null and Timelike Geodesics near the Throats of Phantom Scalar Field Wormholes // Universe. — 2019. — Vol. 6, no. 10. — P. 183.

[9] И.Б. Погребысский. Комм. к Задаче трёх тел Пуанкаре // Пуанкаре А. Избранные труды. Т. 2. — М.:Наука, 1979.

[10] Д.Ф. Селиванов. Курс исчисления конечных разностей. — СПб: тип. Имп. Акад. наук, 1908.

[11] А.А. Марков. Исчисление конечных разностей. — СПб: тип. Имп. Акад. наук, 1889.

[12] Н.Н. Калиткин, П.В. Корякин. Численные методы. Т.2. Методы математической физики. — М.: Академия, 2013.

[13] Э. Хайрер, Г. Ваннер. Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. — М.: Мир, 1999.

[14] А.Г. Свешников, А.Б. Альшин, М.О. Корпусов, Ю.Д. Плетнер. Линейные и нелинейные уравнения соболевского типа. — М.: Физматлит, 2007.

[15] А.А. Самарский, В.А. Галактионов, С.П. Курдюмов, А.П. Михайлов. Режимы с обострением в задачах для квазилинейных параболических уравнений. — М.: Наука, 1987.

[16] К. Бракнер, С. Джорна. Управляемый лазерный синтез. — М.: Атомиздат, 1977.

[17] Л.К. Зарембо, В.А. Красильников. Введение в нелинейную акустику. — М.: Наука, 1966.

[18] О.В. Руденко, С.И. Солуян. Теоретические основы нелинейной акустики. — М.: Наука, 1975.

[19] Ш.У. Галиев. Нелинейные волны в ограниченных сплошных средах. — Киев, Наукова думка, 1988.

[20] C.F. Curtiss, J.O. Hirschfelder. Integration of stiff equations // Proc. Nat. Acad. Sci. — 1952. — Vol. 38. — P. 235–243.

[21] J.C. Butcher. General linear method: a survey // Appl. Num. Math. — 1985. — Vol. 1. — P. 273–284.

[22] H.J. Stetter. Analysis of Discretization Methods for Ordinary Differential Equations. — Springer Tracts in Natural Philosophy. Springer-Verlag Berlin Heidelberg, 2003.

[23] R.C.Aiken (Ed.). Stiff Computation. — Oxford University Press, Oxford, UK, 1985.

[24] J.D. Lambert. Computational methods in ordinary differential equations. — John Wiley and Sons, London, 1973.

[25] L.F. Shampine. Numerical Solution of Ordinary Differential Equations. — Chapman & Hall, New York, 1994.

[26] Ю.В. Ракитский, С.М. Устинов, И.Г. Черноруцкий. Численные методы решения жестких систем. — М.: Наука, 1979.

[27] C.W. Gear. Numerical Initial Value Problems in Ordinary Differential Equations. — Prentice-Hall, Englewood Cliffs, NJ, 1971.

[28] C.W. Gear, D.S. Watanabe. Stability and convergence of variable order multistep methods // SIAM J. Numerical Analysis. — 1974. — Vol. 11. — P. 1044–1058.

[29] C.W. Gear, K.W. Tu. The effects of variable mesh size on the stability of multistep methods // SIAM J. Num. Anal. — 1974. — Vol. 11. — P. 1025–1043.

[30] C.W. Gear. Estimation of errors and derivatives in ordinary differential equations // Proceeedings of IFIP Congress. — 1974. — P. 447–451.

[31] C.W. Gear, K.A. Gallivan. Automatic methods for highly oscillatory differential equations // Lecture notes in Mathematics. — 1982. — Vol. 912. — P. 115–124.

[32] L.F. Shampine, M.W. Reichelt. The Matlab ODE suite // SIAM J. Scientific Computing. — 1997. — Vol. 18, no. 1. — P. 1–22.

[33] J.C. Butcher. Coeffcients for the study of runge-kutta integration processes // J. Austral. Math. Soc. — 1963. — Vol. 3. — P. 185–201.

[34] J.C. Butcher. On Runge-Kutta processes of high order // J. Austral. Math. Soc. — 1964. — Vol. 4. — P. 179–194.

[35] C.R. Cassity. The complete solution of the fifth order Runge-Kutta equations // SIAM J. Numer. Anal. — 1969. — Vol. 6. — P. 432–436.

[36] J.H. Verner. Refuge for Runge-Kutta Pairs. — `http://people.math.sfu.ca/~jverner/`.

[37] P.W. Sharp, J.H. Verner. Generation of high-order interpolants for explicit Runge-Kutta pairs // TOMS. — 1998. — Vol. 24, no. 1. — P. 13–29.

[38] P. Stone. Peter Stone's Maple Worksheets. — `http://www.peterstone.name/Maplepgs`.

[39] Г.М. Хаммуд. Трехмерное семейство 7-шаговых методов Рунге-Кутта порядка 6 // Вычислительные методы и программирование. — 2001. — Vol. 2, no. 2. — P. 71–78.

[40] Е.А. Альшина, Е.М. Закс, Н.Н. Калиткин. Описка в коэффициентах схемы Хаммуда // Вычислительные методы и программирование. — 2007. — Vol. 8, no. 1. — P. 35–37.

[41] Е.А. Альшина, Е.М. Закс, Н.Н. Калиткин. Оптимальные схемы Рунге–Кутты с первого по шестой порядок точности // Ж. вычисл. матем. и матем. физ. — 2008. — Vol. 48, no. 3. — P. 418–429.

[42] S.I. Khashin. A Symbolic-Numeric Approach to the Solution of the Butcher Equations // Canadian Applied Mathematics Quarterly. — 2009. — Vol. 17, no. 1. — P. 555–569.

[43] S.I. Khashin. Butcher algebras for Butcher systems // Numerical Algorithms. — 2012. — Vol. 61, no. 2. — P. 1–11.

[44] С.И. Хашин. Три упрощающих предположения для методов Рунге-Кутта // Вестник ИвГУ. — 2012. — no. 2. — P. 142–150.

[45] S.I. Khashin. Estimating the Error in the Classical Runge–Kutta Methods // Comp. Math. Math. Phys. — 2014. — Vol. 54, no. 5. — P. 767–774.

[46] J.R. Dormand, P.J. Prince. A family of embedded Runge-Kutta formulae // J. Comput. Appl. Math. — 1980. — Vol. 6. — P. 19–26.

[47] J.R. Dormand, P.J. Prince. Runge-Kutta triples // CAMWA. — 1986. — Vol. 12A. — P. 1007–1017.

[48] J. Steven, C. Prentice. Stepsize Selection in Explicit Runge-Kutta Methods for Moderately Stiff Problems // Appl. Math. — 2011. — Vol. 2. — P. 711–717.

[49] H.H. Rosenbrock. Some general implicit processes for the numerical solution of differential equations // The Computer Journal. — 1963. — Vol. 5, no. 4. — P. 329–330.

[50] П.Д. Ширков. Оптимально затухающие схемы с комплексными коэффициентами для жестких систем ОДУ // Матем. моделирование. — 1992. — Vol. 4, no. 8. — P. 47–57.

[51] А.Б. Альшин, Е.А. Альшина, А.Г. Лимонов. Двухстадийные комплексные схемы Розенброка для жестких систем // Ж. вычисл. матем. и матем. физ. — 2009. — Vol. 49, no. 2. — P. 270–287.

[52] С.С. Филиппов. АБС-схемы для жестких систем обыкновенных дифференциальных уравнений // Докл. РАН. — 2004. — Vol. 399, no. 2. — P. 170–172.

[53] М.В. Булатов, А.В. Тыглиян, С.С. Филиппов. Об одном классе одношаговых одностадийных методов для жестких систем обыкновенных дифференциальных уравнений // Ж. вычисл. матем. и матем. физ. — 2011. — Vol. 51, no. 7. — P. 1251–1265.

[54] П.Д. Ширков. Устойчивость ROW методов для неавтономных систем обыкновенных дифференциальных уравнений // Матем. моделирование. — 2012. — Vol. 24, no. 5. — P. 97–111.

[55] А.М. Зубанов, П.Д. Ширков. Численное исследование одношаговых явно-неявных методов, L-эквивалентных жестко точным двухстадийным схемам Рунге–Кутты // Матем. моделирование. — 2012. — Vol. 24, no. 12. — P. 129–136.

[56] Е.А. Новиков, Ю.А. Шитов, Ю.И. Шокин. Одношаговые безытерационные методы решения жестких систем // Докл. АН СССР. — 1988. — Vol. 301, no. 6. — P. 1310–1314.

[57] А.Л. Двинский, Е.А. Новиков. Аппроксимация матрицы Якоби в (m, 3)-методах решения жестких систем // Сиб. журн. вычисл. матем. — 2008. — Vol. 11, no. 3. — P. 283–295.

[58] Е.А. Новиков. L-устойчивый (4,2)-метод четвертого порядка для решения жестких задач // Вестн. СамГУ. Естественнонаучн. сер. — 2011. — Vol. 89, no. 8. — P. 59–68.

[59] И.П. Пошивайло. Жесткие и плохо обусловленные нелинейные модели и методы их расчета : Диссертация ... кандидата физико-математических наук / Пошивайло И.П. ; Институт прикладной математики им. М.В. Келдыша РАН. — 2015.

[60] Н.Н. Калиткин, И.П. Пошивайло. Обратные Ls-устойчивые схемы Рунге-Кутты // ДАН. — 2012. — Vol. 442, no. 2. — P. 175–180.

[61] Н.Н. Калиткин, И.П. Пошивайло. Вычисления с использованием обратных схем Рунге-Кутты // Математическое моделирование. — 2013. — Vol. 25, no. 10. — P. 79–96.

[62] Zhanlav T., Zhuluunbaatar O. New Developments of Newton-Type Iterations for Solving Non-Linear Problems. — Kurs, Moscow, 2022.

[63] R. Anguelov, J.M.-S. Lubuma. Nonstandard finite difference method by nonlocal approximation // Math. Comput. Simul. — 2003. — Vol. 61, no. 3-6. — P. 465–475.

[64] R.E. Mickens (Ed.). Applications of Nonstandard Finite Difference Schemes. — World Scientific, Singapore, 2000.

[65] R.E. Mickens. Nonstandard Finite Difference Models of Differential Equations. — World Scientific, Singapore, 2004.

[66] K.C. Patidar. On the use of nonstandard finite difference methods // J. Difference Eq. Appl. — 2005. — Vol. 11. — P. 735–758.

[67] M. Hochbruck, C. Lubich, H. Selhofer. Exponential integrators for large systems of differential equations // SIAM J. Sci. Comput. — 1998. — Vol. 19, no. 5. — P. 1552–1574.

[68] T. Jahnke, C. Lubich. Error bounds for exponential operator splittings // BIT Numer. Math. — 2000. — Vol. 40, no. 4. — P. 745–744.

[69] J.L. de Lagrange. Theorie des fonctions analytiques, contenant les principes du calcul differentiel, degages de toute consideration d'infiniment petits, d'lvanouissants, de limites et de fluxions, et reduits a l' analyse algebrique des quantites finies. — Paris, 1797, nouv. ed. 1813, Oeuvres Tome 9.

[70] P. Painlevé. Leçons sur la théorie analytique des équations differentielles: : professèes á Stockholm (septembre, octobre, novembre 1895) sur l'invitation de S.M. le roi de Suéde et de Norwège. — Paris, A. Hermann, 1897.

[71] E. Hairer, C. Lubich. Asymptotic expansions of the global error of fixed-stepsize methods // Numer. Math. — 1984. — Vol. 45, no. 3. — P. 354–360.

[72] E. Hairer, C. Lubich, M. Roche. Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations // BIT Numer. Math. — 1988. — Vol. 28, no. 3. — P. 678–700.

[73] E. Hairer, C. Lubich, M. Roche. Error of Rosenbrock methods for stiff problems studied via differential algebraic equations // BIT Numer. Math. — 1989. — Vol. 29, no. 1. — P. 77–90.

[74] C. Lubich. Integration of stiff mechanical systems by Runge-Kutta methods // Zeitschrift fur angewandte Mathematik und Physik ZAMP. — 1993. — Vol. 44, no. 6. — P. 1022–1053.

[75] C. Lubich, K. Nipp, D. Stoffer. Runge–Kutta solutions of stiff differential equations near stationary points // SIAM J. Num. Anal. — 1995. — Vol. 32, no. 4. — P. 1296–1307.

[76] G.J. Cooper. The order of convergence of general linear methods for ordinary differential equations // SIAM J. Numer. Anal. — 1978. — Vol. 15. — P. 643–661.

[77] G.J. Cooper. Error Estimates for General Linear Methods for Ordinary Differential Equations // SIAM J. Numer. Anal. — 1981. — Vol. 18. — P. 65–82.

[78] R. England. Error estimates for Runge-Kutta type solutions to systems of ordinary differential equations // Comput. J. — 1969. — Vol. 18. — P. 166–170.

[79] J.H. Verner. Explicit Runge–Kutta Methods with Estimates of the Local Truncation Error // SIAM J. Numer. Anal. — 1976. — Vol. 15. — P. 772–790.

[80] J.S.C. Prentice. General error propagation in the RKGL method // J. Comput. Appl. Math. — 2009. — Vol. 228. — P. 344–354.

[81] J.S.C. Prentice. Amplification and suppression of round-off error in Runge-Kutta methods // Int. J. Math. Educ. Sci. Technol. — 2009. — Vol. 42, no. 3. — P. 377–385.

[82] J.S.C. Prentice. Error in Runge–Kutta methods // Int. J. Math. Educ. Sci. Technol. — 2009. — Vol. 44, no. 3. — P. 434–442.

[83] P. Kaps, P. Rentrop. Generalized Runge–Kutta methods of order four with stepsize control for stiff ordinary differential equations // Numer. Math. — 1979. — Vol. 33, no. 1. — P. 55–68.

[84] P. Kaps, S.W.H. Poon, T.D. Bui. Rosenbrock methods for Stiff ODEs: A comparison of Richardson extrapolation and embedding technique // Computing. — 1985. — Vol. 34, no. 1. — P. 17–40.

[85] J.R. Cash. Semi-implicit Runge-Kutta procedures with error estimates for the numerical integration of stiff systems of ODEs // JACM. — 1976. — Vol. 23. — P. 455–460.

[86] Y.N.I. Chan, I. Birnbaum, L. Lapidus. Solution of stiff differential equations and the use of imbedding techniques // Industr. Eng. Chemistry Fundamentals. — 1978. — Vol. 17. — P. 133–148.

[87] G. Vanden Berghe, L.Gr. Ixaru, H. DeMeyer. Solution of stiff differential equations and the use of imbedding techniques // Industr. Eng. Chemistry Fundamentals. — 1978. — Vol. 17. — P. 133–148.

[88] A.R. Yaakub, D.J. Evans. A fourth order Runge–Kutta RK(4,4) method with error control // Int. J. Computer Math. — 1999. — Vol. 71, no. 3. — P. 383–411.

[89] F. Iavernaro F. Mazzia. Block-Boundary Value Methods for the solution of ordinary differential equation // SIAM J. Sci. Comput. — 1999. — Vol. 21, no. 1. — P. 323–339.

[90] T.M.H. Chan, J.C. Butcher. Multistep zero approximations for stepsize control // Appl. Numer. Math. — 2000. — Vol. 34. — P. 167–177.

[91] L.F. Shampine, H.A. Watts. Comparing error estimators for Runge–Kutta methods // Math. Comput. — 1971. — Vol. 25. — P. 445–455.

[92] J.C. Butcher, Z. Jackiewicz. A reliable error estimation for DIMSIMs // BIT. — 2001. — Vol. 41. — P. 656–665.

[93] Z. Jackiewicz, J.H. Verner. Derivation and implementation of twostep Runge–Kutta pairs // Appl. Math. — 2002. — Vol. 19. — P. 227–248.

[94] J.C. Butcher, Z. Jackiewicz. A new approach to error estimation for general linear methods // Numer. Math. — 2003. — Vol. 95. — P. 487–502.

[95] J.C. Butcher. Numerical Methods for Ordinary Differential Equations. — Wiley, New York, 2008.

[96] Z. Jackiewicz. General Linear Methods for Ordinary Differential Equations. — Wiley, New York, 2009.

[97] S. Khashin. Estimating the Error in the Classical Runge–Kutta Methods // Comp. Math. Math. Phys. — 2014. — Vol. 54, no. 5. — P. 767–774.

[98] L.F. Shampine. Evaluation of a test set for stiff ODE solvers // ACM Trans. Math. Software. — 1981. — Vol. 7. — P. 409–420.

[99] L.F. Shampine, L.S. Baca. Error estimators for stiff differential equations // J. Comput. Appl. Math. — 1984. — Vol. 11. — P. 197–207.

[100] H. Zedan. A variable order/variable-stepsize Rosenbrock-type algorithm for solving stiff systems of ODE's. Technical Report YCS114. — Department of Computer Science, University of York, York, England, 1989.

[101] A. Jannelli, R. Facio. Adaptive stiff solvers at low accuracy and complexity // J. Comput. Appl. Math. — 2006. — Vol. 191, no. 2. — P. 246–258.

[102] E.A. Celaya, J.J.A. Aguirrezabala, P. Chatzipantelidis. Implementation of an Adaptive BDF2 Formula and Comparison with the MATLAB Ode15s // Proc. Computer Sci. — 2014. — Vol. 29. — P. 1014–1026.

[103] F.T. Krough. Algorithms for changing the step size // SIAM J. Numerical Analysis. — 1973. — Vol. 10. — P. 949–965.

[104] C.H. Hsiao. Numerical solution of stiff differential equations via Harr wavelets // Int. J. Comput. Math. — 2005. — Vol. 82, no. 9. — P. 1117–1123.

[105] J.D. Day. A minimum configuration L-stable fourth-order non-autonomous Rosenbrock method for stiff differential equations // Commun. Appl. Numer. Methods. — 2005. — Vol. 1, no. 6. — P. 293–297.

[106] М.П. Галанин, С.А. Конев. Об одном численном методе решения обыкновенных дифференциальных уравнений // Препринты ИПМ им. М. В. Келдыша. — 2017. — no. 18. — URL: http://keldysh.ru/papers/2017/prep2017_18.pdf.

[107] А.Ф. Филиппов. О корректности и сходимости разностных уравнений // ДАН. — 1955. — Vol. 100, no. 6. — P. 1045–1048; УМН. — 1957. — Vol. 12, no. 1. — P. 245.

[108] R. Courant, K. Friedrichs, H. Lewy. Tiber die partiellen Differentialgleichungen der mathematischen Physik // Math. Annalen. — 1928. — Vol. 100. — P. 32–74.

[109] Р. Курант, К. Фридрихс, Г. Леви. О разностных уравнениях математической физики // УМН. — 1941. — no. 8. — P. 125–160.

[110] В.С. Рябенький, А.Ф. Филлипов. Об устойчивости разностных уравнений. — М.: Государственное изд-во технико-теоретической литературы, 1956.

[111] P. Lax. On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients // Comm. Pure Appl. Math. — 1961. — Vol. 14. — P. 497–520.

[112] Б.Л. Рождественский, Н.Н. Яненко. Системы квазилинейных уравнений. — М.: Наука, 1978.

[113] Р. Рихтмайер, К. Мортон. Разностные методы решения краевых задач. — М.: Мир, 1972.

[114] J. Von Neumann, R.D. Richtmyer. A method for numerical calculation of hydrodynamics shocks // J. Appl. Phys. — 1949. — Vol. 21. — P. 232.

[115] K. Yee. Numerical solution of inital boundary value problems involving Maxwell's equations in isotropic media // IEEE Trans. Antennas. Propag. — 1966. — Vol. 14, no. 3. — P. 302–307.

[116] А.А. Самарский. Введение в теорию разностных схем. — М.: Наука, 1971.

[117] А.А. Самарский. Теория разностных схем. — М.: Наука, 1989.

[118] Е.Г. Дьяконов. Минимизация вычислительной работы. Асимптотически оптимальные алгоритмы для эллиптических задач. — М.: Наука, 1989.

[119] E.G. D'yakonov. Optimization in Solving Elliptic Problems. — CRC Press, Boca Raton, FL, 1996.

[120] Р.П. Федоренко. Введение в вычислительную физику. — М.: Изд-во МФТИ, 1984.

[121] С.К. Годунов, А.В. Забродин, М.Я. Иванов и др. Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.

[122] С.К. Годунов. Разностный метод численного расчета разрывных решений уравнений гидродинамики // Матем. сб. — 1959. — Vol. 47(89), no. 3. — P. 271–306.

[123] А.Н. Тихонов, А.А. Самарский. О сходимости разностных схем в классе разрывных коэффициентов // ДАН. — 1959. — Vol. 8, no. 3. — P. 529–532.

[124] А.А. Самарский, Ю.П. Попов. Разностные методы решения задач газовой динамики. — М.: Наука, 1992.

[125] Е.А. Альшина, Н.Н. Калиткин, П.В. Корякин. Диагностика особенностей точного решения методом сгущения сеток // ДАН. Информатика. — 2005. — Vol. 404, no. 3. — P. 295–299.

[126] Н.Н. Калиткин, П.В. Корякин. Бикомпактные схемы и слоистые среды // ДАН. — 2008. — Vol. 419, no. 6. — P. 744–748.

[127] Н.Н. Калиткин, П.В. Корякин. Одномерные и двумерные бикомпактные схемы в слоистых средах // Матем. моделирование. — 2009. — Vol. 21, no. 8. — P. 44–62.

[128] Е.Ю. Днестровская, Н.Н. Калиткин, И.В. Ритус. Решение уравнений в частных производных схемами с комплексными коэффициентами // Матем. моделирование. — 1991. — Vol. 3, no. 9. — P. 114–127.

[129] А.Н. Тихонов. О зависимости решений дифференциальных уравнений от малого параметра // Матем. сборник. — 1948. — Vol. 22, no. 2. — P. 193–204.

[130] А.Н. Тихонов. О системах дифференциальных уравнений, содержащих параметры // Матем. сборник. — 1950. — Vol. 27, no. 1. — P. 147–156.

[131] А.Н. Тихонов. Системы дифференциальных уравнений, содержащие малые параметры при производных // Матем. сборник. — 1952. — Vol. 31, no. 3. — P. 575–586.

[132] А.Б. Васильева, В.Ф. Бутузов. Асимптотические разложения решений сингулярно возмущенных уравнений. — М.: Наука, 1973.

[133] А.Б. Васильева, В.Ф. Бутузов. Асимптотические методы в теории сингулярных возмущений. — М.: Высшая школа, 1990.

[134] А.Б. Васильева, А.А. Плотников. Асимптотическая теория сингулярно возмущенных задач. — М.: Физический факультет МГУ, 2008.

[135] В.Ф. Бутузов. Асимптотические методы в сингулярно возмущенных задачах. — Ярославль: Изд-во ЯрГУ, 2014.

[136] Н.Н. Нефедов. Метод дифференциальных неравенств для некоторых сингулярно возмущенных задач в частных производных // Дифф. уравнения. — 1995. — Vol. 31, no. 4. — P. 719–722.

[137] V.F. Butuzov, A.B. Vasileva, N.N. Nefedov. Asymptotic theory of contrast structures (review) // Automatics and Remote Control. — 1997. — Vol. 58, no. 7. — P. 1068–1091.

[138] Н.Н. Нефедов. Общая схема асимптотического исследования устойчивых контрастных структур // Нелинейная динамика. — 2010. — Vol. 6, no. 1. — P. 181–186.

[139] V. Volkov, D. Lukyanenko, N. Nefedov. Asymptotic-numerical method for the location and dynamics of internal layers in singular perturbed parabolic problems // Lecture Notes in Computer Science. — 2010. — Vol. 10187. — P. 721–729.

[140] D.V. Lukyanenko, V.T. Volkov, N.N. Nefedov. Dynamically adapted mesh construction for the efficient numerical solution of a singular perturbed reaction-diffusion-advection equation // МАИС. — 2010. — Vol. 24, no. 3. — P. 322–338.

[141] В.Б. Андреев. К анализу одной разностной схемы для уравнения Лапласа в сингулярно возмущённой области // Ж. вычисл. матем. и матем. физ. — 1987. — Vol. 27, no. 10. — P. 1527–1535.

[142] В.Б. Андреев, И.А. Савин. О равномерной по малому параметру сходимости монотонной схемы А. А. Самарского и ее модификации // Ж. вычисл. матем. и матем. физ. — 1995. — Vol. 35, no. 5. — P. 739–752.

[143] Андреев В.Б. Коптева Н.В. Об исследовании разностных схем с аппроксимацией первой производной центральным разностным отношением // Ж. вычисл. матем. и матем. физ. — 1996. — Vol. 36, no. 8. — P. 101–117.

[144] В.Б. Андреев. О сходимости модифицированной монотонной схемы Самарского на гладко сгущающейся сетке // Ж. вычисл. матем. и матем. физ. — 1998. — Vol. 38, no. 8. — P. 1266–1278.

[145] В.Б. Андреев. Функция Грина и априорные оценки решений монотонных трехточечных сингулярно возмущенных разностных схем // Дифференц. уравнения. — 2001. — Vol. 37, no. 7. — P. 880–890.

[146] В.Б. Андреев. Поточечные и весовые априорные оценки решения и его первой производной для сингулярно возмущенного уравнения конвекции-диффузии // Дифференц. уравнения. — 2002. — Vol. 38, no. 7. — P. 918–929.

[147] В.Б. Андреев. Анизотропные оценки функции Грина сингулярно возмущенного двумерного монотонного разностного оператора конвекции-диффузии и их применения // Ж. вычисл. матем. и матем. физ. — 2003. — Vol. 43, no. 4. — P. 546–553.

[148] В.Б. Андреев. О равномерной сходимости на неравномерной сетке классической разностной схемы для одномерного сингулярно возмущенного уравнения реакции-диффузии // Ж. вычисл. матем. и матем. физ. — 2004. — Vol. 44, no. 3. — P. 476–492.

[149] В.Б. Андреев. К теории разностных схем для сингулярно возмущенных уравнений // Дифференц. уравнения. — 2004. — Vol. 40, no. 7. — P. 898–907.

[150] В.Б. Андреев. О точности сеточных аппроксимаций негладких решений сингулярно возмущенного уравнения реакции-диффузии в квадрате // Дифференц. уравнения. — 2006. — Vol. 42, no. 7. — P. 895–906.

[151] В.Б. Андреев. Равномерная сеточная аппроксимация негладких решений смешанной краевой задачи для сингулярно возмущенного уравнения реакции-диффузии в прямоугольнике // Ж. вычисл. матем. и матем. физ. — 2008. — Vol. 48, no. 1. — P. 90–114.

[152] В.Б. Андреев. Оценки в классах Ге?льдера регулярной составляющей решения сингулярно возмущенного уравнения конвекции-диффузии // Ж. вычисл. матем. и матем. физ. — 2017. — Vol. 57, no. 12. — P. 1983–2020.

[153] В.Б. Андреев, И.Г. Белухина. Оценки в классах Гёльдера решения неоднородной задачи Дирихле для сингулярно возмущенного однородного уравнения конвекции-диффузии // Ж. вычисл. матем. и матем. физ. — 2019. — Vol. 59, no. 2. — P. 264–276.

[154] В.Б. Андреев, И.Г. Белухина. Декомпозиция решения двумерного сингулярно возмущенного уравнения конвекции-диффузии с переменными коэффициентами в квадрате; оценки в гёльдеровых нормах // Ж. вычисл. матем. и матем. физ. — 2021. — Vol. 61, no. 2. — P. 206–2016.

[155] Г.И. Шишкин. Сеточные аппроксимации сингулярно возмущенных эллиптических и параболических уравнений. — Екатеринбург; УрО РАН, 1992.

[156] J.J.H. Miller, E. O'Riordan, G.I. Shishkin. Fitted Numerical Methods For Singular Perturbation Problems: Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions. — World Scientific Co. Inc., 2012.

[157] N. Kopteva, E. O'Riordan. Shishkin meshes in the numerical solution Of singularly perturbed differential equations // Internat. J. of Num. Analysis and Modeling. — 2010. — Vol. 7, no. 3. — P. 393–415.

[158] H.-G. Roos, M. Stynes, L. Tobiska. Robust Numerical Methods for Singularly Perturbed Differential Equations. Springer Series in Computational Mathematics. Vol. 24. 2nd ed. — Springer-Verlag, Berlin, 2008.

[159] K. Lipnikov, G. Manzini, M. Shashkov. Mimetic finite difference method // J. Comput. Phys. — 2014. — Vol. 257. — P. 1163–1223.

[160] B. Koren, R. Abgrall, P. Bonchev, J. Frank, B. Perot (eds.). Physics – compatible numerical methods // J. Comput. Phys. — 2014. — Vol. 257, part 2. — P. 1039–1526.

[161] R. de Vogelaere. Methods of integration which preserve the contact transformation property of the Hamiltonian equations. — Dept. Math., Univ. of Notre Dame, Notre Dame, Ind., 1956.

[162] R.D. Ruth. A canonical integration technique // IEEE Trans. Nuclear Science. — 1983. — Vol. NS-30. — P. 2669–2671.

[163] K. Feng. On difference schemes and symplectic geometry // Proceedings of the 5-th Intern. Symposium on differential geometry & differential equations, Aug. 1984, Beijing. — 1985. — P. 42–58.

[164] G.J. Cooper. Stability of Runge–Kutta methods for trajectory problems // IMA J. Numer. Anal. — 1987. — Vol. 7. — P. 1–13.

[165] E. Hairer, C. Lubich, G. Wanner. Geometric numerical integration illustrated by the Stormer-Verlet method // Acta numerica. — 2003. — Vol. 12. — P. 399–450.

[166] F.M. Lasagni. Canonical Runge-Kutta methods // Journal of Applied Mathematics and Physics. — 1988. — Vol. 39. — P. 952–953.

[167] Ю.Б. Сурис. О сохранении симплектической структуры при численном решении гамильтоновых систем. Численное решение обыкновенных дифференциальных уравнений (ред. Филиппов С.С.). — М.: ИПМ. АН СССР, 1956.

[168] Ю.Б. Сурис. Гамильтоновы методы типа Рунге–Кутты и их вариационная трактовка // Матем. моделирование. — 1990. — Vol. 2, no. 4. — P. 78–87.

[169] E. Hairer, C. Lubich, G. Wanner. Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations. — Springer-Verlag Berlin Heidelberg, 2006.

[170] М.Н. Геворкян. Анализ составных симплектических методов и симплектических методов Рунге–Кутта на длительных интервалах времени : Диссертация ... кандидата физико-математических наук / Геворкян М.Н. ; РУДН. — 2013.

[171] Б. Батгэрэл, Э.Г. Никонов, И.В. Пузынин. Процедура вывода явных, неявных и симметричных симплектических схем для численного решения гамильтоновых систем уравнений // Компьютерные исследования и моделирование. — 2016. — Vol. 8, no. 6. — P. 861–871.

[172] E. Faou, E. Hairer, M. Hochbruck, C. Lubich. Geometric Numerical Integration. — Mathematisches Forschungsinstitut Oberwolfach 2016, Report No. 18/2016, 2016.

[173] Nuan Fang Xu, Zi-Chen Deng, Yan Wang, Kai Zhang. A symplectic Runge-Kutta method for analysis of the tethered satellite system // Multidiscipline Modeling in Materials and Structures. — 2016. — Vol. 13, no. 1. DOI: 10.1108/MMMS-11-2016-0060.

[174] J.M. Sanz-Serna. Symplectic Runge–Kutta Schemes for Adjoint Equations, Automatic Differentiation, Optimal Control, and More // SIAM Review. — 2016. — Vol. 58, no. 1. — P. 3–33.

[175] Hong Zhang, Xu Qian, Songhe Song. Novel high-order energy-preserving diagonally implicit Runge–Kutta schemes for nonlinear Hamiltonian ODEs // Applied Mathematics Letters. — 2020. — Vol. 102. — P. 106091.

[176] Hong Zhang, Xu Qian, Jingye Yan, Songhe Song. Highly efficient invariant-conserving explicit Runge-Kutta schemes for the nonlinear Hamiltonian differential equations. — 2019. — 11. DOI: DOI:10.13140/RG.2.2.36232.16643.

[177] L. Brugnano, F. Iavernaro. Line Integral Methods for Conservative Problems. — London New York, CRC Press, 2016.

[178] А.А. Самарский, А.П. Михайлов. Математическое моделирование. Идеи. Методы. Примеры. — М.: Физматлит, 2001.

[179] L.F. Richardson. The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam // Phil. Trans., A. — 1910. — Vol. 210. — P. 307–357.

[180] L.F. Richardson, J.A. Gaunt. The deferred approach to the limit // Phil. Trans. A. — 1927. — Vol. 226. — P. 299–349.

[181] Н.Н. Калиткин. Численные методы. — М.: Наука, 1978.

[182] Л. Гонсалес, М.Д. Малых. О новом пакете для численного решения обыкновенных дифференциальных уравнений в Sage // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Москва, РУДН, 16–20 апреля 2022 г. — Москва : РУДН, 2022.

[183] R. Frank, C.W. Ueberhuber. Iterated defect correction for differential equations. Part I: Theoretical results // Computing. — 1978. — Vol. 20. — P. 207–228.

[184] C.W. Ueberhuber. Implementation of defect correction methods for stiff differential equations // Computing. — 1979. — Vol. 23. — P. 205–232.

[185] A.C. Hindmarsh, G.D. Byrne. EPISODE: An effective package for the integration of systems of ordinary differential equations. Report UCID-30112 (Rev. 1). — Lawrence Livermore Laboratory, Livermore, 1977.

[186] Н.Н. Калиткин, А.Б. Альшин, Е.А. Альшина, Б.В. Рогов. Вычисления на квазиравномерных сетках. — М.: Физматлит, 2005.

[187] Н.Н. Калиткин, Л.В. Кузьмина. Вычисление корней уравнения и определение их кратности // Матем. моделирование. — 2010. — Vol. 22, no. 7. — P. 33–52.

[188] Н.Н. Калиткин, Е.А. Альшина. Численные методы. Т.1. Численный анализ. — М.: Академия, 2013.

[189] Н.Н. Калиткин. Физик от Бога // Экстремальные состояния Льва Альтшулера / Под ред. Б.Л. Альтшулера и В.Е. Фортова. — М.: Физматлит, 2011. — P. 223–229.

[190] Н.Н. Калиткин. Решение задач на собственные значения методом дополненного вектора // ЖВМиМФ. — 1965. — Vol. 5, no. 6. — P. 1107–1115.

[191] А.Ф. Сидоров. Об одном алгоритме расчета оптимальных разностных сеток // Труды МИАН СССР. — 1966. — Рў. 74. — РЎ. 147–151.

[192] Н.Н. Калиткин, Н.О. Кузнецов, С.Л. Панченко. Метод квазиравномерных сеток в бесконечной области // ДАН. — 2000. — Vol. 375, no. 5. — P. 598–601.

[193] Марчук Г.И. Шайдуров В.В. Повышение точности решений разностных схем. — М.: Наука, 1979.

[194] Zienkiewicz O.C. Whiteman J.R. (ed.). The Mathematics of Finite Elements and Applications. — Academic Press, London, 1973.

[195] В.Л. Дербов, В.В. Серов, С.И. Виницкий и др. О решении низкоразмерных краевых задач квантовой механики методом Канторовича – приведение к обыкновенным дифференциальным уравнениям // Известия Саратовского университета: Серия Физика. — 2010. — Vol. 10. — P. 4–17.

[196] А.А. Гусев. Метод конечных элементов высокого порядка точности решения краевых задач для эллиптического уравнения в частных производных // Вестник РУДН. Серия Математика, Информатика, Физика. — 2017. — Vol. 25, no. 3. — P. 217–233.

[197] A.A. Gusev. Algorithm for computing wave functions, reflection and transmission matrices of the multichannel scattering problem in the adiabatic representation using the finite element method // Вестник РУДН. Серия Математика. Информатика. Физика. — 2014. — Vol. 22, no. 2. — P. 93–114.

[198] A.A. Gusev, O. Chuluunbaatar, S.I. Vinitsky et al. POTHEA: A program for computing eigenvalues and eigenfunctions and their first derivatives with

respect to the parameter of the parametric self–adjoined 2D elliptic partial differential equation // Computer Physics Communications. — 2014. — Vol. 185, no. 10. — P. 2636–2654.

[199] A.A. Gusev O. Chuluunbaatar S.I. Vinitsky A.G. Abrashkevich. KANTBP 3.0: new version of a program for computing energy levels, reflection and transmission matrices, and corresponding wave functions in the coupled–channel adiabatic approach // Computer Physics Communications. — 2014. — Vol. 185, no. 10. — P. 3341–3343.

[200] A.A. Gusev O. Chuluunbaatar S.I. Vinitsky et al. Algorithms and programs for solving boundary-value problems for systems of second-order odes with piecewise constant potentials: multichannel scattering and eigenvalue problems // Вестник РУДН: Серия Математика. Информатика. Физика. — 2016. — Рў. 24, в„– 3. — РЎ. 38–52.

[201] A.A. Gusev, S.I. Vinitsky, O. Chuluunbaatar et al. High–Accuracy Finite Element Method: Benchmark Calculations // European Physics Journal – Web of Conferences. — 2018. — Vol. 173. — P. 03010–1–4.

[202] A.A. Gusev, V.P. Gerdt, O. Chuluunbaatar et al. Symbolic–numerical algorithms for solving elliptic boundary–value problems using multivariate simplex lagrange elements // Lecture Notes in Computer Science. — 2018. — Vol. 11077. — P. 197–213.

[203] Н.Н. Калиткин, А.А. Белов. Аналог метода Ричардсона для логарифмически сходящегося счета на установление // ДАН. — 2013. — Vol. 452, no. 3. — P. 261–265.

[204] А.А. Белов, Н.Н. Калиткин. Решение уравнения Фредгольма первого рода сеточным методом с регуляризацией по А.Н. Тихонову // Математическое моделирование. — 2018. — Vol. 30, no. 8. — P. 67–88.

[205] Р.Е. Виноград. Об одном критерии неустойчивости в смысле А. М. Ляпунова решений линейной системы обыкновенных дифференциальных уравнений // ДАН. — 1952. — Vol. 84, no. 2. — P. 201–204.

[206] Б.Ф. Былов, Р.Э. Виноград, Д.М. Гробман, В.В. Немыцкий. Теория показателей Ляпунова и ее приложения к вопросам устойчивости. — М.: Наука, 1966.

[207] К. Деккер, Я. Вервер. Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений (перевод под ред. Самарского). — М.: Мир, 1988.

[208] В.И. Шалашилин, Е.Б. Кузнецов. Метод продолжения решения по параметру и наилучшая параметризация. — М.: Эдиториал УРСС, 1999.

[209] К.Л. Зигель. Лекции по небесной механике. — М.:ИЛ, 1959.

[210] Alfimov G.L., Usero D., Vazquez L. On complex singularities of solutions of the equation $\mathcal{H}u_x - u + u^p = 0$ // J. Phys. A: Math. Gen. — 2000. — Vol. 33. — P. 6707–6720.

[211] Alfimov G.L. On analytic properties of periodic solutions for equation $\mathcal{H}u_x - u + u^p = 0$ // J. Phys. A: Math. Theor. — 2012. — Vol. 45. — P. art. 395205.

[212] Алфимов Г.Л., Лебедев М.Е. О регулярных и сингулярных решениях уравнения $u_{xx} + Q(x)u + P(x)u^3 = 0$ // Уфимский математический журнал. — 2015. — Vol. 7, no. 2. — P. 3–18.

[213] Алфимов Г.Л., Кизин П.П. О решениях задачи Коши для уравнения $u_{xx} + Q(x)u - P(u) = 0$, не имеющих сингулярностей на заданном интервале // Уфимский математический журнал. — 2016. — Vol. 8, no. 4. — P. 24–42.

[214] Alfimov G.L., Kizin P.P., Zezyulin D.A. Gap solitons for the repulsive Gross-Pitaevskii equation with periodic potential: coding and method for

computation // Discr. Cont. Dynam. Syst. Ser. B. — 2017. — Vol. 22. — P. 1207–1229.

[215] Alfimov G.L., Fedotov A.P., Sinelshchikov D.I. Determination of the blow up point for complex nonautonomous ODE with cubic nonlinearity // Physica D: Nonlinear Phenomena. — 2020. — Vol. 402. — P. art.132245.

[216] A. Baddour, M.D. Malykh, A.A. Panin, L.A. Sevastianov. Numerical determination of the singularity order of a system of differential equations // Discrete Continuous Models Appl. Comput. Sci. — 2020. — Vol. 28, no. 1. — P. 17–34.

[217] NIST Digital Library of Mathematical Functions. — `https://dlmf.nist.gov`.

[218] А.И. Толстых. Компактные разностные схемы и их применение в задачах аэрогидродинамики. — М.: Наука, 1990.

[219] Б.В. Рогов, М.Н. Михайловская. О сходимости компактных разностных схем // Математическое моделирование. — 2008. — Vol. 20, no. 1. — P. 99–117.

[220] B.V. Rogov, M.N. Mikhailovskaya. Monotonic bicompact schemes for linear transport equations // Math. Models Comput. Simul. — 2012. — Vol. 4. — P. 92–100.

[221] E.N. Aristova, B. V. Rogov. Boundary conditions implementation in bicompact schemes for the linear transport equation // Math. Models Comput. Simul. — 2013. — Vol. 5. — P. 199–207.

[222] E.N. Aristova, D.F. Baydin, B.V. Rogov. Bicompact scheme for linear inhomogeneous transport equation // Math. Models Comput. Simul. — 2013. — Vol. 5. — P. 586–594.

[223] E.N. Aristova, S.V. Martynenko. Bicompact Rogov schemes for the multidimensional inhomogeneous linear transport equation at large optical depths // Comput. Math. and Math. Phys. — 2013. — Vol. 53. — P. 1499–1511.

[224] M.D. Bragin, B.V. Rogov. Uniqueness of a high-order accurate bicompact scheme for quasilinear hyperbolic equations // Comput. Math. and Math. Phys. — 2014. — Vol. 54. — P. 831–836.

[225] E.N. Aristova, B.V. Rogov, A.V. Chikitkin. Monotonization of a highly accurate bicompact scheme for a stationary multidimensional transport equation // Math. Models Comput. Simul. — 2016. — Vol. 8. — P. 108–117.

[226] E.N. Aristova, B.V. Rogov, A.V. Chikitkin. Optimal monotonization of a high-order accurate bicompact scheme for the nonstationary multidimensional transport equation // Comput. Math. and Math. Phys. — 2016. — Vol. 56. — P. 962–976.

[227] E.N. Aristova, M.I. Stoynov. Bicompact schemes for solving a steady-state transport equation by the quasi-diffusion method // Math. Models Comput. Simul. — 2016. — Vol. 8. — P. 615–624.

[228] Е.Н. Аристова, Н.И. Караваева . Бикомпактные схемы для численного решения модельной задачи нестационарного переноса нейтронов HOLO алгоритмами // Матем. моделирование. — 2021. — Vol. 33, no. 8. — P. 3–26.

[229] Магомедов К.М. Холодов А.С. Сеточно-характеристические численные методы. — М., Наука, 1988.

[230] Е.Б. Кузнецов, С.С. Леонов. Параметризация задачи Коши для систем обыкновенных дифференциальных уравнений с предельными особыми точками // Ж. вычисл. матем. и матем. физ. — 2017. — Vol. 57, no. 6. — P. 934–957.

[231] А.Н. Колмогоров, С.В. Фомин. Элементы теории функций и функционального анализа. — М.: Наука, 1976.

[232] Волчинская М.И. Гольдин В.Я. Калиткин Н.Н. Самарский А.А. Сравнение разностных схем на тестах // Препринт ИПМ им. М.В. Келдыша. — 1972. — no. 44. — P. 1–19.

[233] Ф. Хаусдорф. Теория множеств. — М., Л.: Объединенное научно-техническое издательство НКТП СССР. Главная редакция технико-теоретической литературы, 1937.

[234] В.И.О. Бляшке. Круг и шар. — М.: Наука, 1967.

[235] Лебедев В.И. Явные разностные схемы для решения жестких задач с комплексным или разделимым спектром // Ж. вычисл. матем. и матем. физ. — 2000. — Vol. 40, no. 12. — P. 1801–1812.

[236] А.Б. Васильева, В.Ф. Бутузов, Н.Н. Нефедов. Контрастные структуры в сингулярно возмущенных задачах // Фундаментальная и прикладная математика. — 1998. — Vol. 4, no. 3. — P. 799–851.

[237] Н.Н. Калиткин, И.П. Пошивайло. Решение задачи Коши для жестких систем с гарантированной точностью методом длины дуги // Матем. Моделирование. — 2014. — Рў. 26, в„– 7. — РЎ. 3–18.

[238] Ю.В. Ракитский. О некоторых свойствах решений систем обыкновенных дифференциальных уравнений одношаговыми методами численного интегрирования // Ж. вычисл. матем. и матем. физ. — 2016. — Vol. 1, no. 6. — P. 947–962.

[239] В.Я. Гольдин, Н.Н. Калиткин. Нахождение знакопостоянных решений обыкновенных дифференциальных уравнений // Ж. вычисл. матем. и матем. физ. — 1966. — Vol. 6, no. 1. — P. 162–163.

[240] Н.Н. Калиткин, И.П. Пошивайло. Гарантированная точность при решении задачи Коши методом длины дуги // ДАН. Информатика. — 2013. — Vol. 452, no. 5. — P. 499–502.

[241] Е.Е. Пескова. Моделирование химически реагирующих потоков с использованием вычислительных алгоритмов высокого порядка точности : Диссертация ... кандидата физико-математических наук / Пескова Е.Е. ; Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева. — 2018.

[242] NIST Chemical kinetics database. Standard reference database 17-2Q98. — http://kinetics.nist.gov/kinetics/. — 2011–2016.

[243] Berkeley University of California, Gas Research Institute, GRI-Mech 3.0. — http://www.me.berkeley.edu/gri_mech/. — URL: http://www.me.berkeley.edu/gri_mech/.

[244] Л.Б. Ибрагимова, Г.Д. Смехов, О.П. Шаталов. Сравнительный анализ скоростей химических реакций, описывающих горение водородо-кислородных смесей // Физико-химическая кинетика в газовой динамике. — 2009. — Vol. 8. — P. 1–25, www.chemphys.edu.ru/pdf/2009--06--29--001.pdf.

[245] U. Mass, J. Warnatz. Ignition processes in hydrogen-oxygen mixtures // Combust. Flame. — 1988. — Vol. 74. — P. 53.

[246] А.А. Белов, Н.Н. Калиткин, Л.В. Кузьмина. Моделирование химической кинетики в газах // Математическое моделирование. — 2016. — Vol. 28, no. 8. — P. 46–64.

[247] Н.А. Кудряшов. Аналитическая теория нелинейных дифференциальных уравнений. — Институт компьютеных исследований, М., Ижевск, 2004.

[248] М.А. Лаврентьев, Б.В. Шабат. Методы теории функций комплексного переменного. — М.: Лань, 2002.

[249] А.Г. Свешников, А.Н. Тихонов. Теория функций комплексной переменной. — М.: Физматлит, 2005.

[250] Е.А. Альшина, Н.Н. Калиткин, П.В. Корякин. Диагностика особенностей точного решения при расчетах с контролем точности // Ж. вычисл. матем. и матем. физ. — 2005. — Vol. 45, no. 10. — P. 1837–1847.

[251] E.A. Ayrjan, M.D. Malykh, L.A. Sevastianov. On Difference Schemes Approximating First-Order Differential Equations and Defining a Projective Correspondence Between Layers // J. Math. Sci. — 2019. — Vol. 240. — P. 634–645.

[252] E. Kuester, M. Mohamed, M. Piket-May, C. Holloway. Averaged Transition Conditions for Electromagnetic Fields at a Metafilm // IEEE Trans. Antennas Propag. — 2003. — Vol. 51. — P. 2641.

[253] C. Holloway, M. Mohamed, E. Kuester, A. Dienstfrey. Reflection and Transmission Properties of a Metafilm: With an Application to a Controllable Surface Composed of Resonant Particles // IEEE Trans. Electromagn. Compat. — 2005. — Vol. 47. — P. 853.

[254] D. Morits, C. Simovski. Electromagnetic characterization of planar and bulk metamaterials: A theoretical study // Phys. Rev. B. — 2010. — Vol. 82. — P. 165114.

[255] D. Morits, C. Simovski. Erratum: Electromagnetic characterization of planar and bulk metamaterials: A theoretical study [Phys. Rev. B 82, 165114 (2010) // Phys. Rev. B. — 2012. — Vol. 85. — P. 039901.

[256] C. Holloway, D. Love, E. Kuester, J. Gordon, D. Hill. Use of Generalized Sheet Transition Conditions to Model Guided Waves on Metasurfaces/Metafilms // IEEE Trans. Antennas Propag. — 2012. — Vol. 60. — P. 5173.

[257] А.Ф. Харвей. Техника сверхвысоких частот. — М.: Советское радио, 1965.

[258] Н.Н. Ушаков. Технология элементов вычислительных машин. — М.: Высшая школа, 1976.

[259] Белевцев А.Т. Монтаж радиоаппаратуры и приборов. — М.: Высшая школа, 1975.

[260] В.А. Миличко, А.С. Шалин, И.С. Мухин и др. Солнечная фотовольтаика: современное состояние и тенденции развития // УФН. — 2016. — Vol. 186. — P. 801–852.

[261] М. Борн, Э. Вольф. Основы оптики. — М.: Наука, 1973.

[262] М.Б. Виноградова, О.В. Руденко, А.П. Сухоруков. Теория волн. — М.: Наука, 1979.

[263] Л.А. Вайнштейн. Электромагнитные волны. — М.: Радио и связь, 1988.

[264] С.А. Ахманов, С.Ю. Никитин. Физическая оптика. — М.: Изд-во МГУ; Наука, 2004.

[265] В.И. Денисов. Введение в электродинамику материальных сред. — М.: Изд-во МГУ, 1989.

[266] Я.П. Терлецкий, Ю.П. Рыбаков. Электродинамика. — М.: Высшая школа, 1990.

[267] А.С. Ильинский, В.В. Кравцов, А.Г. Свешников. Математические модели электродинамики. — М.: Высшая школа, 1991.

[268] А.А. Егоров, К.П. Ловецкий, Л.А. Севастьянов, А.А. Хохлов. Многослойные оптические покрытия. — М.: Изд. РУДН, 2014.

[269] G. Bao, P. Li. Maxwell's Equations in Periodic Structures. Applied Mathematical Sciences, vol 208. — Springer, Singapore, 2022. DOI: 10.1007/978-981-16-0061-6_1.

[270] G. Mur. Absorbing Boundary Conditions for the Finite-Difference Approximation of the Time-Domain Electromagnetic-Field Equations // IEEE Trans. Electromagn. Comp., EMC. — 1981. — Vol. 23, no. 4. — P. 377–382.

[271] W.M. Robertson, M.S. May. Surface electromagnetic wave excitation on one-dimensional photonic band-gap arrays // Appl. Phys. Lett. — 1999. — Vol. 74. — P. 1800.

[272] B. Auguié, M.C. Fuertes, P.C. Angelomé et al. Tamm Plasmon Resonance in Mesoporous Multilayers: Toward a Sensing Application // ACS Photonics. — 2014. — Vol. 1, no. 9. — P. 775–780.

[273] B.I. Afinogenov, A.A. Popkova, V.O. Bessonov, A.A. Fedyanin. Measurements of the femtosecond relaxation dynamics of Tamm plasmon-polaritons // Appl. Phys. Lett. — 2016. — Vol. 141. — P. 171107.

[274] R. Brückner, M. Sudzius, S.I. Hintschich et al. Hybrid optical Tamm states in a planar dielectric microcavity // Phys. Rev. B. — 2011. — Vol. 83. — P. 033405.

[275] J. Gessler, V. Baumann, M. Emmerling et al. Electro optical tuning of Tamm-plasmon exciton-polaritons // Appl. Phys. Lett. — 2014. — Vol. 105. — P. 181107.

[276] S.A. Maier. Plasmonics-A Route to Nanoscale Optical Devices // Advanced Materials. — 2001. — Vol. 13, no. 19. — P. 1501–1505.

[277] С.А. Майер. Плазмоника: Теория и приложения / Под ред. С. С. Савинского. — Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2011.

[278] В.В. Климов. Наноплазмоника. — М.: Физматлит, 2009.

[279] Р.Б. Ваганов, Б.З. Каценеленбаум. Основы теории дифракции. — М.: Наука, 1982.

[280] Л.А. Севастьянов, К.П. Ловецкий, А.П. Горобец, О.Н. Бикеев. Методы и алгоритмы решения задач в моделях оптических покрытий. Уч. пособие. — М.: ИПК РУДН, 2008.

[281] К.П. Ловецкий, Л.А. Севастьянов, О.Н. Бикеев и др. Математическое моделирование и методы расчета оптических наноструктур. Уч. пособие. — М.: ИПК РУДН, 2008.

[282] К.П. Ловецкий, Л.А. Севастьянов, М.В. Паукшто, А.А. Жуков. Методы связанных волн расчета оптических покрытий. Уч. пособие. — М.: ИПК РУДН, 2008.

[283] Л.А. Севастьянов, К.П. Ловецкий, Е.Б. Ланеев. Регулярные методы и алгоритмы расчета обратных задач в моделях оптических структур. Уч. пособие. — М.: ИПК РУДН, 2008.

[284] А.А. Егоров, К.П. Ловецкий, Л.А. Севастьянов, А.А. Хохлов. Метод связанных волн расчета дифракционных оптических структур. Уч. пособие. — М.: ИПК РУДН, 2011.

[285] Л.А. Севастьянов, К.П. Ловецкий, А.Л. Севастьянов, А.А. Хохлов. Математическое и компьютерное моделирование оптических наноструктур. Уч. пособие. — М.: ИПК РУДН, 2013.

[286] К.П. Ловецкий, А.Л. Севастьянов, Л.А. Севастьянов, А.А. Хохлов. Устойчивые методы в оптических моделях. Уч. пособие. — М.: ИПК РУДН, 2013.

[287] А.В. Волков, Д.Л. Головашкин, Л.Л. Досколович и др. Методы компьютерной оптики. / Под ред. В.А. Сойфера. — М.: Физматлит, 2003.

[288] А.В. Гаврилов и др. Дифракционная нанофотоника / под ред. В.А. Сойфера. — М.: Физматлит, 2011.

[289] Е.А. Безус, Д.А. Быков, Л.Л. Досколович и др. Дифракционная оптика и нанофотоника / под ред. В.А. Сойфера. — М.: Физматлит, 2014.

[290] В.В. Котляр. Численное решение уравнений максвелла в задачах дифракционной оптики // Компьютерная оптика. — 2006. — Vol. 29. — P. 24–40.

[291] А.А. Егоров. Исследование рассеяния света и определение статистических характеристик нерегулярностей планарных оптических волноводов : Диссертация ... кандидата физико-математических наук / Егоров А.А. ; РУДН. — 1992.

[292] А.А. Егоров. Теория и математическое моделирование рассеяния лазерного излучения в нерегулярном интегрально-оптическом волноводе при наличии шума : Диссертация ... кандидата физико-математических наук / Егоров А.А. ; РУДН. — 2005.

[293] B.I. Afinogenov, V.O. Bessonov, I.V. Soboleva, A.A. Fedyanin. Ultrafast All-Optical Light Control with Tamm Plasmons in Photonic Nanostructures // ACS Photon. — 2019. — Vol. 6, no. 4. — P. 844.

[294] I.R. Capoglu, G.S. Smith. A Total-Field/Scattered-Field Plane-Wave Source for the FDTD Analysis of Layered Media // IEEE Transactions on Antennas and Propagation. — 2008. — Vol. 56, no. 1. — P. 158–169.

[295] U.S. Inan, R.A. Marshall. Numerical electromagnetics. The FDTD method. — Cambridge University Press, Cambridge, 2011.

[296] R. Wang, H. Xia, I.R. Zhang et al. Bloch surface waves confined in one dimension with a single polymeric nanofibre // Nat. Commun. — 2017. — Vol. 8. — P. 14330.

[297] R. Bruckner, A.A. Zakhidov, R. Scholz et al. Phase-locked coherent modes in a patterned metal–organic microcavity // Nature Photon. — 2012. — Vol. 6. — P. 322.

[298] C. Symonds, G. Lheureux, J.P. Hugonin et al. Confined Tamm Plasmon Lasers // Nano Lett. — 2013. — Vol. 13, no. 7. — P. 3179.

[299] R. Badugu, J.R. Lakowicz. Tamm State-Coupled Emission: Effect of Probe Location and Emission Wavelength // J. Phys. Chem. C. — 2014. — Vol. 118, no. 37. — P. 21558.

[300] R. Das, T. Srivastava, R. Jha. Tamm-plasmon and surface-plasmon hybrid-mode based refractometry in photonic bandgap structures // Opt. Lett. — 2014. — Vol. 39, no. 4. — P. 896.

[301] Ж.О. Домбровская, А.Н. Боголюбов. Немонотонность схемы FDTD при моделировании границ раздела между диэлектриками // Ученые записки физического факультета Московского Университета. — 2017. — no. 4. — P. 1740302.

[302] Г.И. Марчук. Методы расщепления. — М.: Наука, 1988.

[303] Z. Meglicki, S.K. Gray, B. Norris. Multigrid FDTD with Chombo // Comp. Phys. Commun. — 2007. — Vol. 176. — P. 109–120.

[304] A. Van Londersele, D. De Zutter, D.V. Ginste. An in-depth stability analysis of nonuniform FDTD combined with novel local implicitization techniques // J. Comp. Phys. — 2017. — Vol. 342. — P. 177–193.

[305] D.S. Balsara, J.J. Simpson. Making a Synthesis of FDTD and DGTD Schemes for Computational Electromagnetics // IEEE J. Multiscale Multiphys. Comp. Techniques. — 2020. — Vol. 5. — P. 99.

[306] F. Hecht. New development in FreeFem++ // Journal of numerical mathematics. — 2012. — Vol. 20, no. 3-4. — P. 251–266.

[307] J. Yvonnet, P. Villon, F. Chinesta. Bubble and Hermite Natural Element Approximations // Lecture Notes on Computational Science and Engineering. — 2007. — P. 283–298.

[308] А.В. Тихонравов. Амплитудно-фазовые свойства спектральных коэффициентов слоистых сред // Ж. вычисл. матем. и матем. физ. — 1985. — Vol. 25, no. 3. — P. 442–450.

[309] Sh.A. Furman, A.V. Tikhonravov. Basics of optics of multilayer systems. — Gif-sur-Yvette, 1992.

[310] A.A Popkova, A.A. Chezhegov, I.V. Soboleva et al. Ultrafast all-optical switching in the presence of Bloch surface waves // JPCS. — Vol. 1461. — 2020. — P. 012134.

[311] M.N. Polyanskiy. Refractive index database. — `https://refractiveindex.info`. — Accessed on 2022-02-13.

[312] L. Gao, F. Lemarchand, M. Lequime. Exploitation of multiple incidences spectrometric measurements for thin film reverse engineering // Opt. Express. — 2012. — Vol. 20, no. 14. — P. 15734–15751.

[313] G.T. Ruck, D.E. Barrick, W.D. Stewart, C.K. Kirchbaum. Radar Cross Section Handbook. Volumes 1 and 2. — Plenum Press, New York, 1970.

[314] K.J. Pascoe. Reflectivity and Transmissivity through Layered Lossy Media: A User-Friendly Approach. Technical Report AFIT/EN-TR-01-07. — Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio, 2001.

[315] D.W. Berreman. Optics in Stratified and Anisotropic Media: 4×4-Matrix Formulation // J. Opt. Soc. Am. — 1972. — Vol. 62, no. 9. — P. 502–510.

[316] IZOVAC Technologies. — https://www.izovac.com/en/, http://www.izovac-coatings.com/en/.

[317] A. Taflove, S.C. Hagness. Computational Electrodynamics: The Finite-Difference Time-Domain Method. — Artech House, London, 2005.

[318] M. Gryga, D. Vala, P. Kolejak et al. One-dimensional photonic crystal for Bloch surface waves and radiation modes-based sensing // Opt. Mater. Express. — 2019. — Vol. 9, no. 10. — P. 4009–4022.

[319] F.I. Baida, M.P. Bernal. Correcting the formalism governing Bloch Surface Waves excited by 3D Gaussian beams // Commun Phys. — 2020. — Vol. 3. — P. 86.

[320] С.Ю. Доброхотов, М.В. Клименко, И.А. Носиков, А.А. Толченников. Вариационный метод расчета лучевых траекторий и фронтов волн цунами, порожденных локализованным источником // ЖВМиМФ. — 2020. — Vol. 60, no. 8. — P. 1439–1448.

[321] И.А. Носиков. Прямой вариационный метод для расчета траекторных характеристик КВ радиотрасс в ионосфере : Диссертация ... кандидата физико-математических наук / Носиков И.А. ; Балтийский федеральный университет им. И. Канта. — 2020.

[322] G.W. Forbes, M.A. Alonso. Using rays better. I. Theory for smoothly varying media // J. Opt. Soc. Am. A. — 2001. — Vol. 18. — P. 1132–1145.

[323] M.A. Alonso, G.W. Forbes. Using rays better. II. Ray families to match prescribed wave fields // J. Opt. Soc. Am. A. — 2001. — Vol. 18. — P. 1146–1159.

[324] M.A. Alonso, G.W. Forbes. Using rays better. III. Error estimates and illustrative applications in smooth media // J. Opt. Soc. Am. A. — 2001. — Vol. 18. — P. 1359–1370.

[325] G.W. Forbes. Using rays better. IV. Refraction and reflection // J. Opt. Soc. Am. A. — 2001. — Vol. 18. — P. 2557–2564.

[326] G.W. Forbes, M.A. Alonso. What on earth is a ray and how can we use them best? // Proc. SPIE. — 1998. — Vol. 3482. — P. 22–31.

[327] M.A. Alonso, G.W. Forbes. Stable aggregates of flexible elements link rays and waves // Optics Express. — 2002. — Vol. 10. — P. 728–739.

[328] G.W. Forbes, M.A. Alonso. The Holy Grail of optical modelling // International Optical Design Conference 2002, Paul Manhart and Jose Sasian, eds., Proc. SPIE. — 2002. — Vol. 4832. — P. 186–197.

[329] M.A. Alonso, G.W. Forbes. Stable aggregates of flexible elements link rays and waves // Nonimaging Optics: Maximum Efficiency Light Transfer VII, Roland Winston ed., Proc. SPIE. — 2002. — Vol. 5185. — P. 125–136.

[330] J. Um, C. Thurber. A fast algorithm for two-point seismic ray tracing // Bull. Seismol. Soc. Am. — 1987. — Vol. 77, no. 3. — P. 972–986.

[331] T.J. Moser, G. Nolet, R. Snieder. Ray bending revisited // Bull. Seismol. Soc. Am. — 1992. — Vol. 88, no. 1. — P. 259–288.

[332] C.J. Coleman. Point-to-point ionospheric ray tracing by a direct variational method // Radio Sci. — 2011. — Vol. 46, no. 5. — P. 1–7.

[333] Ti:Sapphire Coherent MICRA5 laser Manual. — http://lasers.coherent.com/lasers/Micra-5.

[334] M. Berger, R.V. Kohn. A rescaling algorithm for the numerical calculation of blowing-up solutions // Comm. Pure Appl. Math. — 1988. — Vol. 41, no. 6. — P. 841–863.

[335] N.R. Nassif, D. Fayyad, M. Cortas. Sliced-Time Computations with Re-scaling for Blowing-Up Solutions to Initial Value Differential Equations // Computational Science – ICCS 2005. — 2005. — P. 58–65.

[336] J. Filo, A. Hundertmark-Zaušková. A rescaling algorithm for the numerical solution to the porous medium equation in a two-component domain // Comm. Nonlinear Sci. Numer. Simul. — 2016. — Vol. 39. — P. 411–426.

[337] K. Anada, T. Ishiwata, T. Ushijima. A numerical method of estimating blow-up rates for nonlinear evolution equations by using rescaling algorithm // Japan J. Industrial Appl. Math. — 2017. — Vol. 35, no. 1. — P. 33–47.

[338] Nguyen V.T. Numerical analysis of the rescaling method for parabolic problems with blow-up in finite time // Physica D: Nonlinear Phenom. — 2017. — Vol. 339. — P. 49–65.

[339] L.-Y. Chen, N. Goldenfeld, Y. Oono. Renormalization Group Theory for Global Asymptotic Analysis // Phys. Rev. Lett. — 1994. — Vol. 73. — P. 1311–1315.

[340] V.M. Isaia. Numerical simulation of universal finite time behavior for parabolic IVP via geometric renormalization group // Discrete & Continuous Dynam. Syst. – B. — 2017. — Vol. 22, no. 9. — P. 3459–3481.

[341] A. Cangiani, E.H. Georgoulis, I. Kyza, S. Metcalfe. Adaptivity and Blow-Up Detection for Nonlinear Evolution Problems // SIAM J. Sci. Comput. — 2016. — Vol. 38, no. 6. — P. A3833–A3856.

[342] W. Huang, Y. Ren, R.D. Russell. Moving mesh methods based on moving mesh partial differential equations // J. Comp. Phys. — 1994. — Vol. 113. — P. 279–290.

[343] C.J. Budd, Weizhang Huang, R.D. Russell. Moving Mesh Methods for Problems with Blow-Up // SIAM J. Sci. Comput. — 1996. — Vol. 17, no. 2. — P. 305–327.

[344] W.-Q. Ren, X.-P. Wang. An Iterative Grid Redistribution Method for Singular Problems in Multiple Dimensions // J. Comp. Phys. — 2000. — Vol. 159. — P. 246–273.

[345] R.D. Russell, J.F. Williams, X. Xu. MOVCOL4: A Moving Mesh Code for Fourth-Order Time-Dependent Partial Differential Equations // SIAM J. Sci. Comput. — 2007. — Vol. 29. — P. 197–220.

[346] W. Huang, J. Ma, R.D. Russell. A study of moving mesh PDE methods for numerical simulation of blowup in reaction diffusion equations // J. Comp. Phys. — 2008. — Vol. 227, no. 13. — P. 6532–6552.

[347] C.J. Budd, Weizhang Huang, R.D. Russell. Adaptivity with moving grids // Acta Num. — 2009. — Vol. 18. — P. 111–241.

[348] C.J. Budd, R. Carretero-González, R.D. Russell. Precise computations of chemotactic collapse using moving mesh methods // J. Comp. Phys. — 2005. — Vol. 202, no. 2. — P. 463–487.

[349] S. Hamada. On the blow-up problem for the generalized Proudman-Johnson equation // Proceedings of Japan SIAM. — 2009. — Vol. 19. — P. 1–23.

[350] T. Nakagawa. Blowing up of a finite difference solution to $u_t = u_{xx} + u^2$. // Appl. Math. & Optimization. — 1976. — Vol. 2. — P. 337–350.

[351] Y.-G. Chen. Asymptotic behaviours of blowing-up solutions for finite difference analogue of $u_t = u_{xx} + u^{1+\alpha}$ // J. Faculty of Science: Univ. of Tokyo. — 1986. — Vol. 33. — P. 541–574.

[352] A.M. Stuart, M.S. Floater. On the computation of blow-up // European J. Appl. Math. — 1990. — Vol. 1. — P. 47–71.

[353] C. Bandle, H. Brunner. Numerical analysis of semilinear parabolic problems with blow-up solutions // Madrid: Real Academia de Ciencias Exactas, Físicas y Naturales de Madrid. — 1994. — Vol. 88. — P. 203–222.

[354] L. Abia, J.C. López-Marcos, J. Martínez. The Euler method in the numerical integration of reaction-diffusion problems with blow-up // Madrid: Real Academia de Ciencias Exactas, Físicas y Naturales de Madrid. — 1994. — Vol. 88. — P. 203–222.

[355] C. Brandle, F. Quirós, J.D. Rossi. An adaptive numerical method to handle blow-up in a parabolic system // Num. Math. — 2005. — Vol. 102, no. 1. — P. 39–59.

[356] P. Groisman. Totally discrete explicit and semi-implicit Euler methods for a blow-up problem in several space dimensions // Comput. — 2006. — Vol. 76. — P. 325–252.

[357] C.-H. Cho, S. Hamada, H. Okamoto. On the finite differnece approximation for a parabolic blow-up problem // Japan J. Industrial Appl. Math. — 2007. — Vol. 24. — P. 131–160.

[358] C. Brandle, P. Groisman, J.D. Rossi. Fully discrete adaptive methodsfor a blow-up problem // Math. Models Methods Appl. Sci. — 2011. — Vol. 14, no. 10. — P. 1425–1450.

[359] N.R. Nassif, N. Makhoul-Karam, J. Erhel. A globally adaptive explicit numerical method for exploding systems of ordinary differential equations // Appl. Num. Math. — 2013. — Vol. 67. — P. 204–219.

[360] Z.W. Yang, H. Brunner. Blow-up behavior of collocation solutions to Hammerstein-type Volterra integral equations // SIAM J. Numer. Anal. — 2013. — Vol. 51. — P. 2260–2282.

[361] D.W. McLaughlin, G.C. Papanicolaou, C. Sulem, P.L. Sulem. Focusing singularity of the cubic Schrodinger equation // Phys. Rev. A.— 1986.— Vol. 34.— P. 1200–1210.

[362] R. Haynes, C. Turner. A numerical and theoretical study of blow-up for a system of ordinary differential equations using the Sundman transformation // Atlantic Electronic J. Math.— 2007.— Vol. 2, no. 1.— P. 1–13.

[363] C.-H. Cho. On the computation of the numerical blow-up time // Japan J. Industrial Appl. Math.— 2013.— Vol. 30.— P. 331–349.

[364] Cho C.-H. Numerical detection of blow-up: a new sufficient condition for blow-up // Japan J. Industrial Appl. Math.— 2016.— Vol. 33.— P. 81–89.

[365] Cho C.-H. A numerical algorithm for blow-up problems revisitedNumerical Algorithms // Numerical Algorithms.— 2017.— Vol. 75, no. 3.— P. 675–697.

[366] E. Janke, F. Emde, F. Losch. Taffeln horere Functionen.— B.G. Teubbner Verlagsgesellschaft, Stuttgart, 1960.

[367] C.F. Corliss. Integrating ODE's in the complex plane – Pole vaulting // Math. Comp.— 1980.— Vol. 35.— P. 1181–1189.

[368] B. Fornberg, J.A.C. Weideman. A numerical methodology for the Painleve equations // J. Comput. Phys.— 2011.— Vol. 230.— P. 5957–5973.

[369] M. Fasondini, B. Fornberg, J.A.C. Weideman. Methods for the computation of the multivalued Painleve transcendents on their Riemann surfaces // J. Comput. Phys.— 2017.— Vol. 344.— P. 36–50.

[370] I.M. Willers. A new integration algorithm for ordinary differential equations based on continued fraction approximations // Comm. ACM.— 1974.— Vol. 17.— P. 504–508.

[371] A.A. Abramov, L.F. Yukhno. A method for calculating the Painleve transcendents // Appl. Numer. Math. — 2015. — Vol. 93. — P. 262–267.

[372] Айрян Э.А., Малых М.Д., Севастьянов Л.А. О разностных схемах, аппроксимирующих дифференциальные уравнения первого порядка и задающих проективные соответствия между слоями // Записки семинаров ПОМИ. — 2018. — Vol. 468. — P. 202–220.

[373] Баддур А., Малых М.Д., Севастьянов Л.А. О периодических приближенных решениях динамических систем с квадратичной правой частью // Записки семинаров ПОМИ. — 2021. — Vol. 507. — P. 157–172.

[374] W.N. Hansen. Electric Fields Produced by the Propagation of Plane Coherent Electromagnetic Radiation in a Stratified Medium // J. Opt. Soc. Am. — 1968. — Vol. 58, no. 9. — P. 380–390.

[375] N.P.K. Cotter T.W. Preist J.R. Sambles. Scattering-matrix approach to multilayer diffraction // J. Opt. Soc. Am. A. — 1995. — May. — Vol. 12, no. 9. — P. 1097–1103.

[376] А.Г. Свешников, А.В. Тихонравов. Математические методы в задачах анализа и синтеза слоистых сред // Матем. моделирование. — 1989. — Vol. 1, no. 7. — P. 13–38.

[377] В.К. Игнатович. Этюд об одномерном периодическом потенциале // УФН. — 1986. — Vol. 150, no. 1. — P. 145–148.

[378] В.К. Игнатович. Новый метод решения одномерного уравнения Шредингера // ТМФ. — 1991. — Vol. 88, no. 3. — P. 477–480.

[379] J. Chandezon, M.T. Dupuis, G. Cornet, D. Maystre. Multicoated gratings: a differential formalism applicable in the entire optical region // J. Opt. Soc. Am. — 1982. — Vol. 72. — P. 839–846.

[380] L. Li, J. Chandezon. Improvement of the coordinate transformation method for surface-relief gratings with sharp edges // J. Opt. Soc. Am. A. — 1996. — Vol. 13. — P. 2247–2255.

[381] L. Li. Oblique-coordinate-system-based Chandezon method for modeling one-dimensionally periodic multilayer, inhomogeneous, anisotropic gratings // J. Opt. Soc. Am. A. — 1999. — Vol. 16. — P. 2521–2531.

[382] M.G. Moharam, T.K. Gaylord. Diffraction analysis of dielectric surface-relief gratings // J. Opt. Soc. Am. — 1982. — Vol. 72. — P. 1385–1392.

[383] D.C. Dobson. Optimal design for periodic anti-reflective structures for Helmholtz equation // European J. Appl. Math. — 1993. — Vol. 4. — P. 321–340.

[384] M.G. Moharam, E.B. Grann, D.A. Pommet, T.K. Gaylord. Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings // J. Opt. Soc. Am. A. — 1995. — Vol. 12. — P. 1068–1076.

[385] M.G. Moharam, E.B. Grann, D.A. Pommet, T.K. Gaylord. Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach // J. Opt. Soc. Am. A. — 1995. — Vol. 12. — P. 1077–1086.

[386] R.H. Morf. Exponentially convergent and numerically efficient solution of Maxwell's equations for lamellar gratings // J. Opt. Soc. Am. A. — 1995. — Vol. 12. — P. 1043–1056.

[387] L. Li. Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings // J. Opt. Soc. Am. A. — 1996. — Vol. 13. — P. 1024–1035.

[388] G. Bao. Numerical analysis of diffraction by periodic structures: TM polarization // Numer. Math. — 1996. — Vol. 75. — P. 1–16.

[389] P. Lalanne. Improved formulation of the coupled-wave method for two-dimensional gratings // J. Opt. Soc. Am. A. — 1997. — Vol. 14. — P. 1592.

[390] Popov E. (ed.). Gratings: Theory and Numeric Applications. — CNRS, Institut Fresnel UMR, 2014.

[391] Y. Gao, P. Li. Analysis of time-domain scattering by periodic structures // J. Differ. Equ. — 2016. — Vol. 261. — P. 5094–5118.

[392] Y. Gao, P. Li. Electromagnetic scattering for time-domain Maxwell's equations in an unbounded structure // Math. Model. Methods Appl. Sci. — 2017. — Vol. 27. — P. 1843–1870.

[393] Yilei Li, T.F. Heinz. Optical model for thin layers // 2D Materials. — 2018. — Vol. 5, no. 2.

[394] E. Popov, M. Nevirere. Grating theory: new equations in Fourier space leading to fast converging results for TM polarization // J. Opt. Soc. Am. A. — 2000. — Vol. 17. — P. 1073.

[395] L. Li. New formulation of the Fourier modal method for crossed surface-relief gratings // J. Opt. Soc. Am. A. — 1997. — Vol. 14. — P. 2758–2767.

[396] L. Li. Reformulation of the Fourier modal method for surface-relief gratings made with anisotropic materials // J. Modern Opt. — 1998. — Vol. 45. — P. 1313–1334.

[397] L. Li. Justification of matrix truncation in the modal methods of diffraction gratings // J. Optics A: Pure Appl. Opt. — 1999. — Vol. 1. — P. 531–536.

[398] A.V. Tishchenko M. Hamdoun O. Parriaux. Two-dimensional coupled mode equation for grating waveguide excitation by a focused beam // Opt. Quant. Electron. — 2003. — May. — Vol. 35, no. 1. — P. 475–491.

[399] O. Bruno, F. Reitich. Numerical solution of diffraction problems: a method of variation of boundaries // J. Opt. Soc. Am. A. — 1993. — Vol. 10. — P. 1168–1175.

[400] O. Bruno, F. Reitich. Numerical solution of diffractive problems: a method of variation of boundaries II. Dielectric gratings, Padé approximants and singularities // J. Opt. Soc. Am. A. — 1993. — Vol. 10. — P. 2307–2316.

[401] O. Bruno, F. Reitich. Numerical solution of diffractive problems: a method of variation of boundaries III. Doubly-periodic gratings // J. Opt. Soc. Am. A. — 1993. — Vol. 10. — P. 2551–2562.

[402] O. Bruno, F. Reitich. Accurate calculation of diffraction-grating efficiencies // Proceedings of SPIE. — 1993. — Vol. 1919. — P. 236–247.

[403] O. Bruno, F. Reitich. Approximation of analytic functions: a method of enhanced convergence // Math. Comput. — 1994. — Vol. 63. — P. 195–213.

[404] D.P. Nicholls, F. Reitich. Shape deformations in rough surface scattering: cancellations, conditioning, and convergence // J. Opt. Soc. Am. — 2004. — Vol. 21. — P. 590–605.

[405] D.P. Nicholls, F. Reitich. Shape deformations in rough surface scattering: improved algorithms // J. Opt. Soc. Am. — 2004. — Vol. 21. — P. 606–621.

[406] A. Malcolm, D.P. Nicholls. A field expansions method for scattering by periodic multilayered media // J. Acoust. Soc. Am. — 2011. — Vol. 129. — P. 1783–1793.

[407] Y. He, D.P. Nicholls, J. Shen. An efficient and stable spectral method for electromagnetic scattering from a layered periodic structure // J. Comput. Phys. — 2012. — Vol. 231. — P. 3007–3022.

[408] E.M. Purcell, C.R. Pennypacker. Scattering and adsorption of light by nonspherical dielectric grains // Astrophys. J. — 1973. — Vol. 186. — P. 705–714.

[409] B.T. Draine, P.J. Flatau. Discrete-dipole approximation for scattering calculations // J. Opt. Soc. Am. A. — 1994. — Vol. 11. — P. 1491–1499.

[410] T. Wriedt. Generalized Multipole Techniques for Electromagnetic and Light Scatterin. — Elsevier Science, Amsterdam, 1999.

[411] B.T. Draine. Light Scattering by Nonspherical Particles, Theory, Measurements, and Applications. M. I. Mishchenko, J. W. Hovenier, and L. D. Travis, eds. — Academic, 2000. — P. 131–145.

[412] M.A. Yurkin, V.P. Maltsev, A.G. Hoekstra. Convergence of the discrete dipole approximation. II. An extrapolation technique to increase the accuracy // J. Opt. Soc. Am. A. — 2006. — May. — Vol. 23, no. 1. — P. 2592–2601.

[413] M.A. Yurkin, M. Huntermann. Rigorous and fast discrete dipole approximation for particles near a plane interface // J. Phys. Chem. C. — 2015. — May. — Vol. 119, no. 1. — P. 29088–29094.

[414] Yu.A. Eremin A.G. Sveshnikov. Method of discrete sources in scattering theory // Mosc. Univ. Comput. Math. Cybern. — 1992. — May. — Vol. 47, no. 4. — P. 1–10.

[415] Yu.A. Eremin A.G. Sveshnikov. Mathematical models in nanooptics and biophotonics based on the discrete sources method // Comp. Math. Math. Phys. — 2007. — May. — Vol. 47, no. 2. — P. 262–279.

[416] Yu.A. Eremin I.V. Lopushenko. A hybrid scheme of the discrete sources method for analyzing boundary value problems of nano-optics // Mosc. Univ. Comput. Math. Cybern. — 2016. — May. — Vol. 40, no. 1. — P. 1–9.

[417] A.A. Shcherbakov, A.V. Tishchenko. Generalized source method in curvilinear coordinates for 2D grating diffraction simulation // JQSRT. — 2017. — May. — Vol. 187, no. 1. — P. 76–96.

[418] B. Fornberg. A Practical Guide to Pseudospectral Methods. — Cambridge University Press, Cambridge, 1996.

[419] L.N. Trefethen. Spectral Methods in MATLAB. — SIAM, 2000.

[420] G. Bao. Finite element approximation of time harmonic waves in periodic structures // SIAM J. Numer. Anal. — 1995. — Vol. 32. — P. 1155–1169.

[421] G. Bao. Numerical analysis of diffraction by periodic structures: TM polarization // Numer. Math. — 1996. — Vol. 75. — P. 1–16.

[422] G. Bao. Variational approximation of Maxwell's equations in biperiodic structures // SIAM J. Appl. Math. — 1997. — Vol. 57. — P. 364–381.

[423] G. Bao, Y. Cao, H. Yang. Numerical solution of diffraction problems by a least-square finite element method // Math. Method Appl. Sci. — 2000. — Vol. 23. — P. 1073–1092.

[424] G. Bao, H. Yang. A least-squares finite element analysis of diffraction problems // SIAM J. Numer. Anal. — 2000. — Vol. 37. — P. 665–682.

[425] N.H. Lord, A.J. Mulholland. A dual weighted residual method applied to complex periodic gratings // Proc. Roy. Soc. Edinburgh Sect. A. — 2013. — Vol. 469. — P. 20130176.

[426] D.W. Prather. Analysis and synthesis of finite aperiodic diffractive optical elements using rigorous electromagnetic models : Ph.D. thesis / D.W. Prather ; Department of Electrical Engineering, University of Maryland. — 1997.

[427] M. Albani, P. Bernardi. A Numerical Method Based on the Discretization of Maxwell Equations in Integral Form // IEEE Trans. Microwave Theory Techn. — 1974. — May. — Vol. 22, no. 4. — P. 446–450.

[428] A. Christ, H.L. Hartnagel. Three-Dimensional Finite-Difference Method for the Analysis of Microwave-Device Embedding // IEEE Trans. Microwave Theory Techn. — 1987. — May. — Vol. 35, no. 8. — P. 688–696.

[429] K. Beilenhoff, W. Heinrich, H.L. Hartnagel. Improved finite-difference formulation in frequency domain for three-dimensional scattering problems // IEEE Trans. Microwave Theory Techniques. — 1992. — May. — Vol. 40, no. 3. — P. 540–546.

[430] S.C. Brenner, L.R. Scott. The Mathematical Theory of Finite Element Methods. — Springer, New York, 1994.

[431] P.G. Ciarlet. The Finite Element Method for Elliptic Problems. — North-Holland, Amsterdam, 1978.

[432] J.-M. Jin. The Finite Element Methods in Electromagnetics. — Wiley, New York, 2002.

[433] P. Monk. Finite Element Methods for Maxwell's Equations. — Oxford University Press, Oxford, 2003.

[434] J.S. Hesthaven. High-Order Accurate Methods in Time-Domain Computational Electromagnetics: A Review // Advances in imaging and electron physics. — 2003. — Vol. 127. — P. 59–123.

[435] J.C. Nédélec. Mixed Finite Elements in R^3 // Numer. Math. — 1980. — Vol. 35. — P. 315–341.

[436] R.C. Kirby, A. Logg, A.R. Terrel. Automated Solution of Differential Equations by the Finite Element Method. — 2012. — Vol. 84. — P. 91–116.

[437] J.S. Hesthaven, D. Gottlieb. Stable spectral methods for conservation laws on triangles with unstructured grids // Comput. Methods Appl. Mech. Engin. — 1999. — Vol. 175. — P. 361–381.

[438] J.S. Hesthaven. Spectral penalty methods // Appl. Numer. Math. — 2000. — Vol. 33. — P. 23–41.

[439] J.S. Hesthaven, C. H. Teng. Stable spectral methods on tetrahedral elements // SIAM J. Sci. Comput. — 2000. — Vol. 21. — P. 2352–2380.

[440] K. Dridi, J.S. Hesthaven, A. Ditkowski. Staircase-free finite-difference time-domain formulation for general materials in complex geometries // IEEE Trans. Antennas Propag. — 2001. — Vol. 49, no. 5. — P. 749–756.

[441] J.S. Hesthaven, T. Warburton. High-order nodal methods on unstructured grids. I. Time-domain solution of Maxwell's equations // J. Comput. Phys. — 2002. — Vol. 181. — P. 1–34.

[442] S. Piperno, L. Fezoui. A discontinuous Galerkin FVTD method for 3D Maxwell Equations. — 2003.

[443] J.S. Hesthaven, T. Warburton. High order nodal discontinuous Galerkin Methods for the Maxwell eigenvalue problem // Phil. Trans. R. Soc. Lond. A. — 2004. — Vol. 362. — P. 493–524.

[444] J.A. Cox, D. Dobson. An integral equation method for biperiodic diffraction structures // International Conference on the Applications and Theory of Periodic Structures (Proc. SPIE). — 1991. — Vol. 1545. — P. 106–113.

[445] W. Lu, Y.Y. Lu. High order integral equation method for diffraction gratings // J. Opt. Soc. Am. A. — 2012. — Vol. 29. — P. 734–740.

[446] E. Popov, B. Bozhkov, D. Maystre, J. Hoose. Integral method for echelles covered with lossless or absorbing thin dielectric layers // Appl. Opt. — 1999. — Vol. 38. — P. 47–55.

[447] D.W. Prather, M.S. Mirotznik, J.N. Mait. Boundary integral methods applied to the analysis of diffractive optical elements // J. Opt. Soc. Am. A. — 1997. — Vol. 14. — P. 34–45.

[448] Y. Wu, Y.Y. Lu. Boundary integral equation Neumann-to-Dirichlet map method for gratings in conical diffraction // J. Opt. Soc. Am. A. — 2011. — Vol. 28. — P. 1191–1196.

[449] V. Yachin, K. Yasumoto. Method of integral functionals for electromagnetic wave scattering from a double-periodic magnetodielectric layer // J. Opt. Soc. Am. A. — 2007. — Vol. 24. — P. 3606–3618.

[450] A. Rathsfeld, G. Schmidt, B.H. Kleemann. On a fast integral equation method for diffraction gratings // Commun. Comput. Phys. — 2006. — Vol. 1. — P. 984–1009.

[451] C. Johnson. Numerical solution of partial differential equations by the finite element method. — Cambridge University Press, 1987.

[452] I. Babuška, W.C. Rheinboldt. Error estimates for adaptive finite element computations // SIAM J. Numer. Anal. — 1978. — Vol. 15. — P. 736–754.

[453] W. Dörfler. A convergent adaptive algorithm for Poisson's equations // SIAM J. Numer. Anal. — 1996. — Vol. 33. — P. 1106–1124.

[454] R. Verfürth. A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques. — Teubner, Stuttgart, 1996.

[455] Z. Chen, S. Dai. Adaptive Galerkin methods with error control for a dynamic Ginzburg-Landau model in superconductivity // SIAM J. Numer. Anal. — 2001. — Vol. 38. — P. 1961–1985.

[456] J.M. Cascon, C. Kreuzer, R.H. Nochetto, K.G. Siebert. Quasi-optimal convergence rate for an adaptive finite element method // SIAM J. Numer. Anal. — 2008. — Vol. 46. — P. 2524–2550.

[457] P. Morin, R.H. Nochetto, K.G. Siebert. Convergence of adaptive finite element methods // SIAM Rev. — 2002. — Vol. 44. — P. 631–658.

[458] Z. Chen, S. Dai. On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients // SIAM J. Sci. Comput. — 2002. — Vol. 24. — P. 443–462.

[459] K. Mekchay, R.H. Nochetto. Convergence of adaptive finite element methods for general second order linear elliptic PDEs // SIAM J. Numer. Anal. — 2005. — Vol. 43. — P. 1803–1827.

[460] R. Stevenson. Optimality of a standard adaptive finite element method // Found. Comput. Math. — 2007. — Vol. 7. — P. 245–269.

[461] P. Monk. A posteriori error indicators for Maxwell's equations // J. Comput. Appl. Math. — 1998. — Vol. 100. — P. 173–190.

[462] P. Monk, E. Süli. The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals // SIAM J. Numer. Anal. — 1998. — Vol. 36. — P. 251–274.

[463] P. Morin, R.H. Nochetto, K.G. Siebert. Data oscillation and convergence of adaptive FEM // SIAM J. Numer. Anal. — 2000. — Vol. 38. — P. 466–488.

[464] H.M. Liddel. Computer-aided techniques for the design of multilayer filters. — Bristol-Hilger, 1981.

[465] H.K. Pulker. Coating on glass. — Elsevier Amsterdam-Oxford-New-York-Tokio, 1984.

[466] H.A. Macleod. Thin-film optical filters. — Bristol-Hilger, 1986.

[467] A. Thelen. Design of optical interference coating. — N.Y.: McGraw-Hill, 1989.

[468] М.А. Канибологпский, Ю.С. Уржумцев. Оптимальное проектирование конструкций. — Новосибирск: Наука, 1989.

[469] A. Goncharsky, A. Goncharsky, D. Melnik et al. Nanooptical elements for visual verification // Sci. Rep. — 2011. — Vol. 1. — P. 2426.

[470] A. Goncharsky, S. Durlevich. DOE for the formation of the effect of switching between two images when an element is turned by 180 degrees // Sci. Rep. — 2020. — Vol. 10. — P. 10606.

[471] A. Goncharsky, S. Durlevich. High-resolution computer-generated hologram for creating 2D images with kinematic effects of motion // J. Opt. — 2020. — Vol. 22. — P. 115702.

[472] Г.Д. Бабе, Е.Л.Гусев. Математические методы синтеза слоистых структур. — Новосибирск: Наука, 1987.

[473] Е.Л. Гусев. Качественные закономерности взаимосвязи параметров в оптимальных структурах в задачах оптимального синтеза неоднородных структур из дискретного набора материалов при волновых воздействиях // Доклады РАН. — 1996. — Vol. 346, no. 3. — P. 324–326.

[474] Е.Л. Гусев. Качественные закономерности структуры оптимальных решений в задачах оптимального синтеза многослойных конструкций при воздействии упругих волн // Доклады РАН. — 1998. — Vol. 368, no. 1. — P. 53–56.

[475] Е.Л. Гусев. Качественные закономерности взаимосвязи параметров в слоистых структурах, реализующих предельные возможности // Акуст. журн. — 1999. — Vol. 45, no. 4. — P. 502–506.

[476] Е.Л. Гусев. Свойство внутренней симметрии вструктуре оптимальных слоистых конструкций // Акуст. журн. — 2001. — Vol. 47, no. 1. — P. 43–48.

[477] В.Н. Бакулин, Е.Л. Гусев, В.Г. Марков. Методы оптимального проектирования конструкций из традиционных и композиционных материалов приволновых и статических воздействиях // Инженерно-физический журн. — 2001. — no. 6. — P. 53–57.

[478] В.Н. Бакулин, Е.Л. Гусев, В.Г. Марков. Об оптимальном синтезе слоистых неоднородных структур // Акуст. журн. — 2002. — Vol. 48, no. 3. — P. 325–330.

[479] В.Н. Бакулин, Е.Л. Гусев, В.Г. Марков, А.И. Емельянов. Оптимальное проектирование и численный расчет конструкций с применением композиционных и традиционных материалов // Математическое моделирование. — 2002. — Vol. 14, no. 9. — P. 71–77.

[480] Е.Л. Гусев. Предельные возможности слоистых структур при воздействии акустических волн // Известия РАН. Механика твердого тела. — 2003. — no. 1. — P. 173–179.

[481] Л.А. Севастьянов, К.П. Ловецкий, Е.Б. Ланеев, О.Н. Бикеев. Алгоритмы вычислительного эксперимента для проектирования оптических наноструктур. — М.: РУДН, 2008.

[482] А.А. Белов, А.Н. Боголюбов, Ж.О. Домбровская, С.О. Жбанников. Сверхбыстрый метод расчёта одномерных задач фотоники // Физические основы приборостроения. — 2020. — Vol. 9, no. 2. — P. 2–9.

[483] K.L. Shlager, J.B. Schneider. A selective survey of the finite-difference time-domain literature // IEEE Antennas Propagat. Mag. — 1995. — Vol. 37, no. 4. — P. 39–57.

[484] D.M. Sullivan. Electromagnetic simulation using the FDTD method. — IEEE Press, 2000.

[485] A. Taflove, S.G. Johnson, A. Oskooi. Advances in FDTD Computational Electromagnetics: Photonics and Nanotechnology. — Artech House, London, 2013.

[486] Д.Л. Головашкин. Дифракция H-волны на двумерной идеально проводящей решетке // Матем. моделирование. — 2005. — Vol. 17, no. 4. — P. 53–61.

[487] J.P. Berenger. Three-Dimensional Perfectly Matched Layer for the Absorption of Electromagnetic Waves // J. Comput. Phys. — 1996. — Vol. 127, no. 2. — P. 363–367.

[488] Н.Н. Калиткин. Улучшенная факторизация параболических схем // ДАН. — 2005. — Vol. 402, no. 4. — P. 467–471.

[489] А.А. Самарский, Е.С. Николаев. Методы решения сеточных уравнений. — М.: Наука, 1978.

[490] С.И. Кудрявцев А.Н. Трашкеев. Формализм двух потенциалов для численного решения уравнений Максвелла // Ж. вычисл. матем. и матем. физ. — 2013. — Vol. 153, no. 11. — P. 1823–1834.

[491] К.В. Вязников, В.Ф. Тишкин, А.П. Фаворский. Построение монотонных разностных схем повышенного порядка аппроксимации для систем уравнений гиперболического типа // Матем. моделирование. — 1989. — Vol. 1, no. 5. — P. 95–120.

[492] A. Harten, S. Osher. Uniformly High-Order Accurate Nonoscillatory Schemes. I // SIAM J. Numer. Anal. — 1987. — Vol. 24, no. 2. — P. 279–309.

[493] X.-D. Liu, S. Osher, T. Chan. Weighted Essentially Non-oscillatory Schemes // J. Comput. Phys. — 1994. — Vol. 115, no. 1. — P. 200–212.

[494] A.C. Cangellaris, D.B Wright. Analysis of the numerical error caused by the stair-stepped approximation of a conducting boundary in FDTD simulations of electromagnetic phenomena // IEEE Trans. Antennas. Propag. — 1991. — Vol. 39, no. 10. — P. 1518–1525.

[495] G. R. Werner, J. R. Cary. A stable FDTD algorithm for non-diagonal, anisotropic dielectrics // J. Comp. Phys. — 2007. — Vol. 226. — P. 1085–1101.

[496] A.F. Oskooi, C. Kottke, S.G. Johnson. Accurate finite-difference time-domain simulation of anisotropic media by subpixel smoothing // Opt. Lett. — 2009. — Vol. 34, no. 18. — P. 2778–2780.

[497] C.A. Bauer, G.R. Werner, J.R. Cary. A second-order 3D electromagnetic algorithm for curved interfaces between anisotropic dielectrics on a Yee mesh // J. Comput. Phys. — 2011. — Vol. 230, no. 5. — P. 2060–2075.

[498] G.R. Werner, C.A. Bauer, J.R. Cary. A more accurate, stable, FDTD algorithm for electromagnetics in anisotropic dielectrics // J. Comput. Phys. — 2013. — Vol. 255. — P. 436–455.

[499] T. Hirono, Y. Shibata, W.W. Lui, S. Seki, Y. Yoshikuni. The second-order condition for the dielectric interface orthogonal to the Yee-lattice axis in the FDTD scheme // IEEE Microwave Guided Wave Lett. — 2000. — Vol. 10, no. 9. — P. 359–361.

[500] K.P. Hwang, A.C. Cangellaris. Effective permittivities for second-order accurate FDTD equations at dielectric interfaces // IEEE Microw. Wireless Compon. Lett. — 2001. — Vol. 11, no. 4. — P. 158–160.

[501] R.B. Armenta, C.D. Sarris. A Second-Order Domain-Decomposition Method for Modeling Material Interfaces in Finite-Difference Discretizations // Proceedings of the IEEE MTT-S International. — 2012. — P. 502–505. — URL: https://doi.org/10.1109/MWSYM.2012.6258425.

[502] D.M. Sullivan. Frequency-dependent FDTD methods using Z transforms // IEEE Trans. Antennas Propagat. — 1992. — Vol. 40, no. 10. — P. 1223–1230.

[503] D.M. Sullivan. Z-Transform theory and the FDTD method // IEEE Trans. Antennas Propagat. — 1996. — Vol. 44, no. 1. — P. 28–34.

[504] K. Abdijalilov, H. Grebel. Z-transform theory and FDTD stability // IEEE Trans. Antennas Propagat. — 2004. — Vol. 52, no. 11. — P. 2950–2954.

[505] D.F. Kelley, T.J. Destan, R.J. Luebbers. Debye Function Expansions of Complex Permittivity Using a Hybrid Particle Swarm-Least Squares Optimization Approach // IEEE Trans. Antennas Propagat. — 2007. — Vol. 55, no. 7. — P. 1999–2005.

[506] Z. Lin, Y. Fang, J. Hu, C. Zhang. On the FDTD Formulations for Modeling Wideband Lorentzian Media // IEEE Trans. Antennas Propagat. — 2011. — Vol. 59, no. 4. — P. 1338–1346.

[507] X.T. Dong, N.V. Venkatarayalu, B. Guo, W.Y. Yin, Y.B. Gan. General formulation of unconditionally stable ADI-FDTD method in linear dispersive media // IEEE Trans. Microwave Theory Techn. — 2004. — Vol. 52, no. 1. — P. 170–174.

[508] J.L. Young, R.O. Nelson. A summary and systematic analysis of FDTD algorithms for linearly dispersive media // Antennas Propag. Magazine IEEE. — 2001. — Vol. 43, no. 1. — P. 61–126.

[509] H. Cai, X. Hu, B. Xiong, M.S. Zhdanov. Finite-element time-domain modeling of electromagnetic data in general dispersive medium using adaptive Pade

series // Antennas Propag. Magazine IEEE. — 2017. — Vol. 109, no. 12. — P. 194–205.

[510] Z. Lin, C. Zhang, P. Ou, Y. Jia, L. Feng. A Generally Optimized FDTD Model for Simulating Arbitrary Dispersion Based on the Maclaurin Series Expansion // J. Lightwave Technol. — 2010. — Vol. 28, no. 19. — P. 2843–2850.

[511] W.H. Weedon, C.M. Rappaport. A general method for FDTD modeling of wave propagation in arbitrary frequency-dispersive media // IEEE Trans. Antennas Propagat. — 1997. — Vol. 45, no. 3. — P. 401–410.

[512] J.A. Pereda, A. Vegas, A. Prieto. FDTD modeling of wave propagation in dispersive media by using the Mobius transformation technique // IEEE Trans. Microwave Theory Techn. — 2002. — Vol. 50, no. 7. — P. 1689–1695.

[513] J.C. Bolomey, C. Durix, D. Lesselier. Time domain integral equation approch for inhomogeneous and dispersive slab problems // IEEE Trans. Antennas Propagat. — 1978. — Vol. AP-26, no. 9. — P. 658–667.

[514] R. Luebbers, F.P. Hunsberger, K.S. Kunz et al. A frequency-dependent finite-difference time-domain formulation for dispersive materials // IEEE Trans. Electromagn. Compatibility. — 1990. — Vol. 32, no. 3. — P. 222–227.

[515] R. Luebbers, F.P. Hunsberger, K.S. Kunz. A frequency-dependent finite-difference time-domain formulation for transient propagation in plasmas // IEEE Trans. Antennas Propagat. — 1991. — Vol. 39, no. 1. — P. 29–43.

[516] R. Luebbers, F.P. Hunsberger. FDTD for Nth-order dispersive media // IEEE Trans. Antennas Propagat. — 1992. — Vol. 40, no. 11. — P. 1297–1301.

[517] I. Giannakis A. Giannopoulos. A Novel Piecewise Linear Recursive Convolution Approach for Dispersive Media Using the Finite-Difference Time-

Domain Method // IEEE Trans. Antennas Propagat. — 2014. — Vol. 62, no. 5. — P. 2669–2678.

[518] J.L. Young. Propagation in linear dispersive media: Finite difference time-domain methodologies // IEEE Trans. Antennas Propagat. — 1995. — Vol. 43, no. 4. — P. 422–426.

[519] T. Kashiwa N. Yoshida I. Fukai. A treatment by the finite-difference time-domain method of the dispersive characteristics associated with orientation polarization // Inst. Electron. Inform. Communicat. Eng. Trans. — 1990. — Vol. E73, no. 8. — P. 1326–1328.

[520] T. Kashiwa I. Fukai. A treatment by the FD-TD method for the dispersive characteristics associated with electronic polarization // Microwave Opt. Tech. Lett. — 1990. — Vol. 3, no. 6. — P. 203–205.

[521] O.P. Gandhi. A frequency-dependent finite-difference time-domain formulation for general dispersive media // IEEE Trans. Microwave Theory Tech. — 1993. — Vol. 41, no. 4. — P. 658–665.

[522] R.M. Joseph S.C. Hagness A. Taflove. Direct time integration of Maxwell's equations in linear dispersive media with absorption for scattering and propagation of femtosecond electromagnetic pulses // Opt. Lett. — 1991. — Vol. 16, no. 18. — P. 1412–1414.

[523] F. Maradei. A frequency-dependent WETD formulation for dispersive materials // IEEE Trans. Magnet. — 2001. — Vol. 37, no. 5. — P. 3303–3306.

[524] G. Kobidze J. Gao B. Shanker E. Michielssen. A fast time domain integral equation based scheme for analyzing scattering from dispersive objects // IEEE Trans. Magnet. — 2005. — Vol. 53, no. 3. — P. 1215–1226.

[525] X. Zhuansun X. Ma. Integral-Based Exponential Time Differencing Algorithms for General Dispersive Media and the CFS-PML // IEEE Trans. Antennas Propagat. — 2012. — Vol. 60, no. 7. — P. 3257–3264.

[526] P.G. Petropoulos. Stability and phase error analysis of FDTD in dispersive dielectrics // IEEE Trans. Antennas Propagat. — 1994. — Vol. 42, no. 1. — P. 62–69.

[527] J.L. Young, A. Kittichartphayak, Y.M. Kwok, D. Sullivan. On the dispersion errors related to $(FD)^2$/TD type schemes // IEEE Trans. Microwave Theory Techn. — 1995. — Vol. 43, no. 8. — P. 1902–1910.

[528] Ж.О. Домбровская, А.Н. Боголюбов. Повышение точности одномерной схемы Йе методом сгущения сеток // Изв. РАН. Сер. физ. — 2017. — Vol. 81, no. 1. — P. 117–120.

[529] S. G. Garcia T. W. Lee S. C. Hagness. On the accuracy of the ADI-FDTD method // IEEE Antennas Wirel. Propag. Lett. — 2002. — Vol. 1, no. 1. — P. 31–34.

[530] B. Fornberg. Some Numerical Techniques for Maxwell's Equations in Different Types of Geometries // Topics in Computational Wave Propagation. Lecture Notes in Computational Science and Engineering, vol 31. — 2003. — P. 265–299.

[531] B. Fornberg, J. Zuev, J. Lee. Stability and accuracy of time-extrapolated ADI-FDTD methods for solving wave equations // J. Comput. Appl. Math. — 2007. — Vol. 200. — P. 178–192.

Приложение 1. Обзор методов для задач Коши с сингулярностями

П.1.1. Обнаружение ближайшей сингулярности

В литературе рассмотрены многие подобные задачи. Наиболее хорошо изучены задачи, сводящиеся к параболическому уравнению с нелинейной правой частью [334–362] и в некоторых случаях – с нелинейным пространственным оператором [15].

Для расчета таких задач обычно используют разностные схемы. Однако вблизи сингулярности решение разностных уравнений быстро теряет точность. Чтобы преодолеть эту трудность, предлагались разные подходы.

В [334–338] был предложен метод масштабирования. По мере приближения к сингулярности шаги пространственной и временной сеток уменьшались в соответствии с соотношением подобия. Ожидается, что такое уменьшение шагов должно уменьшить погрешность численного решения. Однако лишь уменьшение шага по времени проводится естественным образом. Уменьшение шага по пространству требует введения новых узлов пространственной сетки. Начальные значения в новых узлах надо находить пространственной интерполяцией, что вносит дополнительную погрешность. Поскольку такое сгущение сеток проводится неоднократно по мере приближения к сингулярности, вклад погрешности интерполяции может стать значительным.

Близко к описанному подходу примыкают другие геометрические методы, основанные на построении нормализующих групп (то есть на других преобразованиях подобия) [339, 340]. Такие подходы могут более адекватно подобрать правило уменьшения шагов для различных конкретных задач. Но в остальном они сохраняют недостатки, присущие методу масштабирования.

В [341] были предложены адаптивные сетки, основанные на апостериорных мажорантных оценках погрешности. Такие оценки достаточно трудно строить, и обычно они имеют громоздкую форму. Не всегда удается получить оценки,

применимые при произвольных шагах сетки. Нередко теоретические оценки доказаны лишь при достаточно малых шагах сетки. Все это затрудняет их практическое применение. Сам момент разрушения определяется как момент «развала» мажорантной оценки, что нельзя считать строгим критерием.

Большое количество работ посвящено методу движущихся сеток [342–347] и его применению в различных задачах [343, 348, 349]. Смысл его в том, что пространственная сетка должна адаптироваться к решению. Тем самым положения узлов сетки должны непрерывно смещаться с течением времени. Для нахождения кривых координатных линий такой сетки предлагаются различные критерии, не имеющие строго обоснования. При этом для нахождения линий сетки возникает вспомогательное уравнение в частных производных, решение которого представляет отдельную трудность: нередко расчетные координатные линии оказываются осциллирующими. Это бессмысленно с геометрической точки зрения, поэтому приходится вводить дополнительное сглаживание, также основанное на нестрогих соображениях. В итоге делается столько приближений, что оценить их суммарный вклад не представляется возможным. Поэтому вопрос о точности расчета сингулярности остается открытым.

В работах [350–360] очередной шаг по времени выбирался по пространственной норме решения на последнем временном слое (обычно обратно пропорционально величине этой нормы). Зачастую удается доказать, что сумма этих уменьшающихся шагов стремится к конечной величине, которая интерпретируется как момент разрушения решения. Этот подход представляется разумным. Однако подобные доказательства справедливы лишь на фиксированных пространственных сетках. Поэтому для полной реализации этого подхода необходимо, во-первых, на каждой пространственной сетке проводить расчет большого числа шагов по времени. Во-вторых, необходимо проводить такие расчеты с многократным сгущением пространственных сеток. Это делает данный подход весьма трудоемким. При этом остается нерешенным вопрос, какую погреш-

ность вносит конечность самых подробных сеток и конечность числа шагов по времени.

Для некоторых конкретных задач удавалось найти преобразование, которое переводит решение в несингулярное [361, 362]. Однако такие случаи весьма редки, а общего алгоритма построения таких преобразований не предложено.

Предлагались и другие частные подходы. Упомянем, например, работы [363–365]. В них вводится формальный критерий «численного разрушения», для которого исследуются достаточные условия сходимости. Однако проверить применимость этих критериев в реальных практических задачах обычно не удается, что отмечают сами авторы этих работ [364].

Описанные выше методы не универсальны и предназначены для конкретных классов задач. Однако существует достаточно общий подход, применимый практически ко всем типам задач [181]. Пусть исходная задача описывается нестационарным уравнением в частных производных. Вводится пространственная сетка, и пространственные операторы заменяются разностными соотношениями. Тогда задача сводится к задаче Коши для системы большого числа обыкновенных дифференциальных уравнений. Интегрирование такой системы по времени проводится с помощью некоторой независимой схемы, не коррелированной с пространственной схемой. Неограниченное нарастание численного решения такой системы при приближении к некоторому моменту времени трактуется как сингулярность численного решения. Затем одновременно проводится сгущение шагов по пространству и времени. Число обыкновенных дифференциальных уравнений при этом соответственно увеличивается. Снова проводится численный расчет и определяется его момент сингулярности. Такое сгущение сеток проводится столько раз, сколько позволяют вычислительные мощности.

Последовательность полученных решений и моментов их сингулярностей, согласно теоремам Рябенького-Филлипова, стремится к точному решению и его моменту разрушения. Этот метод был развит в работах [14, 125, 250] и позволил единообразно решить ряд различных задач.

П.1.2. Последовательность сингулярностей

Существует второй класс задач, в котором решение не разрушается, а имеет множественные сингулярности. В таких задачах требуется найти ряд последовательно расположенных сингулярностей. Подобные задачи часто встречаются в теории специальных функций (эллиптические функции, гамма-функция и т.д.).

Для составления таблиц специальных функций [366] и для стандартных программ прямого расчета [217] широко применяют численные методы. Стандартные схемы (например, схемы Рунге-Кутты) позволяют рассчитывать гладкие участки решения с хорошей точностью. Однако вблизи сингулярности ошибка таких схем катастрофически нарастает. Прямое продолжение решения за полюс, как правило, невозможно. Поэтому решение продолжается за полюс какими-то искусственными приемами. Прохождение ряда полюсов представляет еще большую проблему и требует разработки специальных процедур.

В литературе описаны методы, основанные на Паде-аппроксимации [367–369] и на приближении решения цепными дробями [370]. Абрамов и Юхно предложили специальную замену неизвестной функции, переводящую решение в несингулярное, см. [371] и библиографию там. Однако эти методы применимы только для расчета трансцендент Пенлеве, для которых имеется много априорной информации. Кроме того, коэффициенты Паде-аппроксимации вычисляются по коэффициентам ряда Тейлора, а для нахождения последних нужно решать исходную задачу некоторой разностной схемой. Возникающие при этом проблемы описаны ранее.

В работах Малых и Севастьянова был разработан способ продолжения за полюс для уравнений с квадратичной нелинейностью [372, 373].

Приложение 2. Обзор разностных методов для системы уравнений Максвелла

П.2.1. Методы в частотной области

К этому классу методов относят методы решения стационарной задачи. Перечислим их.

П.2.1.1. Матричные методы

Методы матриц рассеяния (S-matrix methods) [268,315,374,375] применимы для задачи п. 5.1.4 и задачи п. 5.1.3, если пластины являются однородными и поверхностные токи отсутствуют $\sigma^{\text{surf}} = 0$.

Матричные методы основаны на том, что для каждой однородной пластины выписывается точное решение в виде прямой и обратной волны с неизвестными амплитудами. Затем эти решения сшиваются на границах раздела согласно условиям сопряжения (5.9). Это проводит к системе алгебраических уравнений относительно амплитуд, которая решается методом Гаусса или каким-либо итерационным методом. Таким образом, при сравнительно низкой трудоемкости этот метод дает точное решение задачи указанных задач. Разработаны обобщения этого метода на случай наклонного падения и анизотропного материала пластин.

Метод Берремана широко применялся в школе Севастьянова в Российском университете дружбы народов для расчета спектральных характеристик оптических покрытий [268, 280–286]. Большое количество прикладных расчетов выполнил Ловецкий с помощью разработанной им программы MorphoVision.

Неоднородные среды. В работе Свешникова и Тихонравова [376] приведено обобщение этого метода на случай неоднородных пластин. Для каждой пластины решение строится в квадратурах, которые вычисляются каким-либо численным методом, затем применяется описанная выше процедура сшивания таких решений. По существу, этот метод является сеточным. Его точность опре-

деляется точностью используемых численных квадратур. Аналогичный подход развивался Игнатовичем для задач квантово-механического рассеяния частицы на периодическом потенциале [377, 378].

Преобразование координат. В работах [379–381] предложен подход, который позволяет применять матричные методы к расчету рассеяния монохроматических волн дифракционными решетками. Основная идея заключается во введении преобразования координат, такого, что в новых координатах рассеиватель становится плоско-параллельным.

П.2.1.2. Модовые методы

Метод связанных волн. Для задач дифракции монохроматического излучения на периодических прозрачных объектах широко применяется метод связанных волн (rigorous coupled wave analysis, RCWA) [382–393]. В областях пространства, в которых показатель преломления является постоянным, поле представляют в виде линейной комбинации плоских волн (т.е. ряда Фурье по пространственным гармоникам) с неизвестными коэффициентами. Далее такие разложения подставляют в условия сопряжения на границах раздела сред. Это приводит к системе уравнений относительно коэффициентов разложения. Для непериодического ограниченного рассеивателя вместо ряда Фурье по дискретным гармоникам применяют интеграл Фурье [390, 394].

Этот метод активно развивается в коллективе под руководством Сойфера в Институте систем обработки изображений РАН [287–289]. Этим коллективом разработана двумерная реализация метода RCWA. С ее помощью были решены следующие задачи [290]:

- оперативный расчет трехмерных полей дифракции на элементах микрооптики (например, бинарная микролинза),
- задачи дифракции на двумерных решетках из магнитных и анизотропных материалов (например, синтез бинарных антиотражающих структур с различными типами отверстий),

• задачи синтеза многопорядковых дифракционных решеток и ряд других.

Метод связанных волн активно также развивается коллективом под руководством Севастьянова в Российском университете дружбы народов [268, 280–286]. Этот подход применялся для решения как прямых, так и обратных задач. Так, Егоровым была решена задача восстановления статистических характеристик нерегулярностей волновода по диаграмме направленности рассеянного излучения [291, 292].

Фурье-методы. Близко к методу RCWA примыкают так называемые модальные Фурье-методы [387, 395–398]. Последние отличаются от методов типа RCWA улучшенной сходимостью. Нередко в литературе эти подходы объединяют под общим названием метода связанных волн.

Граничные возмущения. Метод RCWA непосредственно применим для структур с прямоугольной геометрией: набору плоско-параллельных пластин, прямоугольной дифракционной решетке и т.п. Чтобы применять его к более сложным задачам, например, к набору пластин с переменной толщиной, используют теорию возмущений. Этот подход был реализован в работах [399–403] применительно к задаче дифракции на структуре из двух слоев. Он получил название метода граничных возмущений. Флуктуация положения границы раздела δ считается малым возмущением. Показано, что электромагнитные поля являются аналитическими функциями этого возмущения. Это позволяет построить фурье-коэффициенты электромагнитных полей в виде ряда по степеням δ и применить методы типа RCWA.

Разложение преобразованного поля. Этот метод представляет собой дальнейшее развитие метода граничных возмущений. В работах [404–407] он реализован для периодической структуры, период которой может содержать произвольное количество слоев с нерегулярными границами.

П.2.1.3. Параметрические методы

Данная группа методов основана на разложении решения по некоторой системе функций. Коэффициенты разложения определяются из условия наилучшего приближения входных данных с помощью этой системы функций. Число слагаемых, требуемых для достижения заданной точности, зависит от того, насколько удачно выбран базис в конкретной задаче.

Разложение по собственным модам. В качестве базиса можно взять собственные моды рассеивателя. Тогда задача сводится к разложению падающей волны по этим собственным модам. Этот подход применяют, например, в задачах расчета поперечных пространственных мод нерегулярных волноводов. Собственные моды находят, решая векторные уравнения Гельмгольца для полей $\mathbf{E}$ и $\mathbf{H}$. Для этого как правило используют разностные методы, приводящие к решению алгебраической задачи на собственные значения. Этот подход развивает Котляр в Институте систем обработки изображений РАН [290].

Согласованные синусоидальные волны. В качестве базиса можно выбрать систему синусоидальных мод. Такой подход получил название метода согласованных синусоидальных мод [289]. Задача нахождения пространственных мод оптических волноводов сводится к решению нелинейной матричной задачи на собственные значения.

Коллектив Сойфера широко применял этот метод, решая алгебраическую задачу на собственные значения методом Крылова. Были разработаны скалярная и векторная формулировки этого метода. Метод синусоидальных волн использовался, например, для расчета низших мод для модели круглого оптического волокна. Было исследовано пространственное распределение энергии по сечению волновода.

Мультипольная аппроксимация. Метод дискретной мультипольной аппроксимации применяют для расчета дифракции на ограниченных рассеивателях сложной формы [408–411]. Рассеиватель разбивают на элементы неболь-

шого размера. Отклик каждого такого элемента представляют с помощью разложения по мультиполям. В простейшем случае ограничиваются только дипольным слагаемым. Сходимость понимают в смысле измельчения разбиения, а число учитываемых мультиполей обычно считают фиксированным [412]. В работе [413] построено обобщение этого подхода на случай, когда ограниченный рассеиватель расположен под неограниченной плоской подложкой.

К этой группе методов близко примыкает метод дискретных источников, развиваемый Свешниковым и его учениками [414–416] в МГУ им. М.В. Ломоносова, см. также [417].

Псевдоспектральные методы. В этой группе методов в качестве базиса выбирают полиномы, например, Чебышева или Лежандра [418, 419].

П.2.1.4. Сеточные методы

Уравнение Гельмгольца. В монохроматическом случае уравнения Максвелла сводятся к уравнению Гельмгольца для потенциалов электромагнитного поля либо непосредственно для полей **E** и **H**. Для них составляют разностные схемы с помощью метода конечных элементов (Finite-Element Frequency Domain, FETD) [420–426] в постановке Ритца либо Галеркина. Применяют также метод разностной аппроксимации (Finite-Difference Frequency Domain, FDFD) [295]. Известны также реализации метода конечных элементов непосредственно для уравнений Максвелла [427–433] Эти подходы представляются наиболее универсальными: задача может иметь сложную геометрию, сетки могут быть неструктурированными, а материал рассеивателя – пространственно неоднородным.

Основной проблемой этих методов является потеря точности вблизи границ раздела из-за нарушения условий сопряжения [434]. Поэтому используют конечные элементы специального вида. Конечные элементы Неделéка [435] обеспечивают тангенциальную непрерывность сеточной функции. Элементы Равьяра-Томаса (см., например, [436] и цитированную литературу) дают нормальную

непрерывность сеточной функции. Однако эти условия являются частными случаями физических граничных условий (а именно, граничное условие для вектора напряженности магнитного поля **H** включает разрыв, пропорциональный величине поверхностных токов). В многочисленных реализациях разрывного метода Галеркина [437–443], условия на границе раздела также нарушаются. Вместо них вводятся фиктивные потоки между ячейками. Эти потоки рассматриваются как степени свободы. В результате в расчете возникают нефизичные решения [434].

Интегральные методы. Задача дифракции может рассматриваться в интегральной постановке. Она приводит к интегральному уравнению Фредгольма второго рода. В квантовой механике аналогичное уравнение называется уравнением Липпмана-Швингера. Для задач дифракции этот подход удобен, поскольку в нем не требуется постановка дополнительных условий излучения для описания ухода излучения на бесконечность.

Для решения таких задач применяют методы конечных и граничных элементов [444–450], а также метод разностной аппроксимации [289]. Эти методы представляются наиболее универсальными. Трудности, с которыми они сталкиваются, описаны выше. Методы конечных и граничных элементов широко использовались группой под руководством Котляра в Институте систем обработки изображений РАН [290].

Для быстрого решения системы разностных уравнений широко применяют итеративный алгоритм, основанный на быстром преобразовании Фурье [289]. Последнее сильно ограничивает выбор сеток: сетки должны быть прямоугольными, шаги по всем переменным – постоянными, а число шагов должно равняться 2^N, где N – целое число. Достоинством алгоритма является исключительно быстрая сходимость – экспоненциальная: погрешность очередной итерации примерно равна квадрату предыдущей.

Описанный подход широко применялся коллективом под руководством Сойфера [289]. В частности, сотрудниками этого коллектива была решена задача

расчета сил, действующих на диэлектрический микроцилиндр со стороны сходящегося пучка. Было показано, что при определенных параметрах пучка и микроцилиндра возможен оптический захват, т.е. существует точка вблизи перетяжки пучка, в которой сумма сил, действующих на цилиндр, равна нулю. Коллектив Сойфера также проводил расчеты дифракции плоской волны на цилиндрических дифракционных микролинзах, а также на диэлектрических микроцилиндрах с произвольным сечением (например, линза Лунеберга).

Оценки сходимости и адаптированные конечные элементы. Как известно, уравнение Гельмгольца является эллиптическим. Согласно лемме Сеа, метод конечных элементов для эллиптических задач имеет первый порядок точности [451]. Эта оценка является априорной. На основе подобных априорных оценок в литературе предложен [452–460] ряд методов адаптации конечно-элементных сеток к решению. Например, характерный размер ячеек сетки в данной области пространства должен быть тем меньше, чем больше оценка погрешности в этой области пространства. Такие сетки называются априорно адаптированными.

Другой подход заключается в адаптации сетки в ходе самого расчета [461–463]. Вычисление проводится не на единственной сетке, а на наборе сгущающихся сеток. При этом очередная сетка перестраивается по решению, полученному на предыдущей сетке. Например, характерный размер ячеек сетки в данной области пространства должен быть тем меньше, чем больше производные решения в этой области пространства. Такие сетки называют апостериорно адаптированными.

П.2.1.5. Синтез оптических покрытий

Большое практическое значение имеют задачи синтеза тонкопленочных оптических покрытий [283, 464–468]. В таких задачах требуется найти такие толщины слоев пленок и материалы, из которых они изготовлены, чтобы спектр отражения или прохождения имел заданный вид. Изготовление таких покры-

тий – это сложная технологическая задача. Поэтому одновременно с задачей синтеза возникает задача контроля процесса изготовления.

Обратная задача синтеза сводится к многократному решению прямой задачи, т.е. вычислению спектра рассеивателя при заданных толщинах и материалах. В оптических задачах (см. п. 5.1.4) чаще всего применяют простые и эффективные матричные методы.

Выдающиеся результаты в этом направлении были получены в Научно-исследовательском вычислительном центре МГУ им. М.В. Ломоносова коллективом под руководством Тихонравова (см., например, [308, 309, 376]). В частности, А.А. Гончарским были разработаны дифракционные оптические элементы, формирующие двумерные и трехмерные изображения, динамические изображения, защитные оптические элементы для визуального и автоматического контроля и др. (см., например, [469–471]).

Ряд важных результатов был получен Гусевым (Институт физико-технических проблем Севера СО РАН) [472–480]. Он исследовал разрешимость задачи синтеза оптических покрытий, то есть существование оптимальных покрытий, максимизирующих или минимизирующих отражение электромагнитной волны. Гусевым был установлен ряд новых качественных закономерностей структуры оптимальных многослойных покрытий, получены оценки для оптимального числа слоев конструкции, получена система рекуррентных соотношений, позволяющая априори до проведения численных расчетов выделить именно те материалы допустимого набора, которые могут входить в оптимальную конструкцию и ряд других результатов. На основе полученных результатов Гусев сделал вывод о том, что многослойные покрытия из существующих природных материалов позволяют получить произвольный наперед заданный характер спектра. Также выявленные закономерности позволили существенно сузить множество допустимых вариантов оптимизируемых покрытий.

Систематическое изложение технологии вычислительного эксперимента в задачах проектирования наноструктур дано Севастьяновым, Ланеевым и соав-

торами в работе [481]. Обзор методов решения обратных задач дан в монографии [283].

П.2.1.6. Выбор метода

Из приведенного обзора следует, что для оптических задач п. 5.1.4, в которых слои являются пространственно однородными, наиболее эффективны матричные методы. Если материалы слоев неоднородны, то необходимо использовать сеточно-матричный метод Свешникова-Тихонравова. Для задач п. 5.1.2 и 5.1.3 при отсутствии поверхностных токов наиболее эффективны, по-видимому, методы конечных и граничных элементов. Для задач п. 5.1.2 и 5.1.3, в которых присутствуют поверхностные токи (то есть решение испытывает сильный разрыв), методы отсутствуют.

П.2.2. Методы во временной области

К этому группе относят методы решения нестационарной задачи.

П.2.2.1. Нестационарный матричный метод

Классические матричные методы, описанные выше, применимы только к стационарным задачам. Автор диссертации построил [482] обобщение матричного метода Берремана на нестационарную задачу п. 5.1.7. Он применим в предположении, что имеет место частотная дисперсия, а пространственная пренебрежимо мала (см. п. 5.1.5). Этот подход обеспечивает сверхбыструю сходимость: погрешность зависит от шага не степенным образом $O(h^p)$, как в классических разностных методах, а экспоненциальным $\sim \exp(-h^{-1})$. Такая скорость убывания погрешности кардинально быстрее степенной. Этот результат не выносится на защиту.

П.2.2.2. Методы конечных разностей и конечных элементов

Эти методы широко применяются для нестационарных уравнений Максвелла в дифференциальной форме [295,483–485]. В зарубежной литературе за ними закрепились названия Finite-Difference Time Domain (FDTD) и Finite-Element Time Domain (FETD) соответственно. Практически все методы этих классов сводятся к схеме «с перешагиваниеим» [115], в которой электрическое и магнитное поля относятся к целым и полуцелым временны́м слоям. Схемы подобного типа впервые появились в газодинамике и были исследованы Курантом [108, 109], см. также [124].

Граничные условия. Традиционно падение волны задают с помощью фиктивного источника поля (метод разделения полного и рассеянного полей TF/SF) [485]. Головашкиным в коллективе Сойфера [486] было предложено «прозрачное» граничное условие, собой некоторую модификацию метода TF/SF.

Уход волны на бесконечность описывают с помощью условий Мура [270] либо с помощью поглощающих стенок (т.н. идеально согласованного слоя, Perfectly Matched Layer, PML) [487], либо с помощью условий Мура [270]. Условия Мура более физичны, однако они имеют лишь первый порядок точности. При использовании PML возникают переотражения. Это вынуждает включать в PML-границу много слоев, что увеличивает трудоемкость расчета, в противном случае точность ухудшается [295].

Неявные схемы. Явные схемы метода FDTD являются лишь условно устойчивыми, поэтому для них необходимо соблюдать выполнение условия Куранта. Для неявных схем этого условия нет, однако они оказываются гораздо более трудоемкими. В литературе описан ряд схем этого класса [295]. Одна из таких схем была построена Головашкиным [289]. В одномерном случае работоспособны чисто неявная схема и схема Кранка-Николсон. В многомерном случае на каждом временном слое нужно решать линейную систему с сильно разреженной матрицей огромной размерности. Для экономичного решения таких задач

проводят их факторизацию, то есть расщепляют многомерную задачу на произведение одномерных. Для прямоугольной геометрии наиболее эффективной оказывается эволюционная факторизация, предложенная Калиткиным [488].

В тех случаях, когда факторизовать систему не удается, для решения линейной системы применяют методы градиентного спуска и распараллеливают программу [289, 290].

Используя методы типа FDTD, коллектив Сойфера решил ряд важных прикладных задач [289, 290]. Среди них моделирование рассеяния на субволновом антиотражающем микрорельефе, прохождения излучения через дифракционные микролинзы и др. Также были выполнены расчеты, в которых исследовалось влияние технологических погрешностей на работу одиночного дифракционного элемента.

Численная дисперсия. Схемы методов FDTD и FETD обладают численной дисперсией [295,485]. Расчетная фазовая скорость отличается от точной, т.е. численное решение отстает по фазе от точного. Известно также [485], что численная дисперсия анизотропна. Иначе говоря, скорости распространения плоских волн по оси координатной сетки или по диагонали неодинаковы. Искривленный волновой фронт (например, цилиндрический или сферический) при распространении «самопроизвольно» деформируется.

Отсюда следует, что схема Йе не может быть строго консервативной. В каждой ячейке дисбаланс отличен от нуля, но его величина стремится к нулю при уменьшении шага сетки. Такие схемы называют почти консервативными. Как известно [489], наличие дисбаланса приводит к ухудшению количественной точности.

Немонотонность. Схемы методов FDTD и FETD имеют теоретический порядок точности, равный 2 и более. Поэтому, согласно теореме Годунова [122], они являются немонотонными. В [301] исследовано влияние немонотонности в случае схемы Йе. В этой работе показано, что для неоднородных сред в численном решении возникают нефизичные пилообразные осцилляции, которые могут

быть сильны. В результате точность резко ухудшается, особенно вблизи границ раздела сред. Фактический порядок точности оказывается лишь первым [301].

П.2.2.3. Уравнения Максвелла в интегральной форме

Для нестационарных уравнений Максвелла в интегральной форме применяют метод конечных объемов во временной области (Finite-Volume Time Domain, FVTD) [295]. В нем также строят схему «с перешагиванием»: аналогично классической схеме Йе ячейки магнитного поля смещены относительно ячеек электрического поля на половину шага по времени.

П.2.2.4. Газодинамические методы

Кудрявцев и Трашкеев предложили нетривиальный подход, который заключается в сведении электродинамической задачи к системе уравнений газодинамического типа с помощью формализма потенциалов [490]. Это позволяет использовать хорошо разработанные почти монотонные разностные схемы высокого порядка точности. Среди них, например, методы, предложенные в работах Тишкина и Фаворского [491], схемы классов Essential non-oscillatory (ENO) [492], Weighted essential non-oscillatory (WENO) [493] и др.

П.2.2.5. Границы раздела

Схемы методов FDTD, FETD, FVTD не являются компактными, поскольку ячейка Йе занимает 1.5 шага по пространству (ячейки магнитного поля сдвинуты на полшага относительно узлов электрического поля). Поэтому как бы ни были выбраны узлы пространственной сетки, граница раздела сред (и, значит, разрыв поля) попадает внутрь ячейки. В этом случае условие аппроксимации схемы Йе нарушается. Заметим, что схема обладает некоторым «запасом прочности» и в случае слабых разрывов обеспечивает сходимость (но медленную). Поэтому на практике погрешность расчета слоистых диэлектрических и маг-

нитных сред оказывается большой [301, 494]. Поэтому для расчетов задач в слоистых средах в литературе предлагались различные специальные подходы. Перечислим их.

Усреднение показателя преломления. В [495–498] предлагалось усреднение показателя преломления в ячейке, внутрь которой попадает граница раздела. В [499,500] разрыв показателя преломления сглаживали, то есть заменяли границу раздела средой с непрерывно изменяющимся показателем преломления. Авторы работы [501] проводили искусственную сшивку решений до и после границы раздела с помощью фиктивных ячеек. Однако перечисленные подходы вносят физическую погрешность в решение. Так, нетрудно заметить, что усреднение показателя преломления вносит погрешность $O(1)$ вблизи границы раздела. Известно также, что размывание или усреднение границ раздела может сильно ослаблять расчетное отражение. Пример такого расчета приведен в п. 6.2.3.

Схемы с выделением особенности. В литературе рассматривались [317] схемы FDTD с явным выделением особенности. Они применялись к задачам с филаментными токами. Такой подход работоспособен, если имеется одна граница раздела. Наличие уже двух границ раздела приводит к многократным переотражениям, и схемы с явным выделением особенностей становятся весьма громоздкими.

Специальные конечные элементы. Как и в стационарном случае, в методах FETD можно применять элементы Неделéка [435] или Равьяра-Томаса [436]. Напомним, что первые обеспечивают тангенциальную непрерывность сеточной функции, а вторые – нормальную непрерывность. Однако эти методы не позволяют учесть физические условия сопряжения типа (5.2) в общем виде, т.е. при наличии ненулевых поверхностных токов.

П.2.2.6. Дисперсия материалов

Многие материалы (например, Si, Ta и их оксиды, Ge и др.) имеют частотную дисперсию. Чтобы это учесть, в уравнение для вектора электрической индукции добавляют интегральный член, содержащий свертку напряженности электрического поля и диэлектрической восприимчивости. Система уравнений Максвелла становится интегро-дифференциальной. Такие задачи намного более трудны, чем чисто дифференциальные.

Для некоторых законов дисперсии (полиномиальный, отношение двух полиномов, модели Дебая и Лоренца) эту свертку удается вычислить точно [502–512]. Однако реальные законы дисперсии приходится приближать комбинацией таких модельных законов. Другой подход заключается в прямом сеточном вычислении этой свертки, для которой фактически записывается некоторая явная схема [513–525]. Оба способа учета дисперсии в методах FDTD, FETD, FVTD вносят погрешность в результаты расчетов, причем она нередко оказывается весьма существенной [526, 527].

П.2.2.7. Оценки точности

Для нестационарых разностных схем применим метод сгущения сеток и оценки точности по методу Ричардсона-Калиткина. Применительно к схеме Йе эта процедура описана в [528]. Подчеркнем, что для применения данного метода необходимо проводить глобальное сгущение сетки, причем одновременно по всем переменным.

В [529] исследована аппроксимация и выписан остаточный член схемы Йе. Однако авторы не приводят ни априорных, ни апостериорных оценок точности.

В работах [495,498] анонсируется применение экстраполяции по методу Ричардсона. Однако результат экстраполяции, полученный в этих работах, не соответствует теоретически ожидаемому (см. [528]). Причина этого в том, что из-за усреднения показателя преломления локальная ошибка вблизи границы разде-

ла есть $O(1)$. В этом случае оценки точности по Ричардсону неправомерны, а экстраполяция может ухудшать точность.

В литературе описан также ряд подходов, в которых сгущение проводилось только по пространственным переменным. Однако при записи нестационарных разностных схем присутствует также погрешность, вызванная аппроксимацией производных по времени.

В работах [530, 531] использовались оценки точности и экстраполяция по Ричардсону, но сгущение проводилось только для сетки по времени, причем не на каждом шаге ее шаге. Эта процедура использовалась для построения адаптивной сетки по времени. Однако она не дает представления о вкладе аппроксимации пространственных производных в погрешность.

П.2.2.8. Выбор метода

Нестационарные задачи намного труднее стационарных. Для них не удается построить простых аналитических методов, аналогичных методу матриц рассеяния, методу связанных волн и т.д. Из приведенного обзора следует, что для нестационарых задач п. 5.1.5 – 5.1.7 при отсутствии поверхностных токов наиболее работоспособны методы конечных разностей и конечных элементов во временной области. Эти методы имеют ряд существенных недостатков, однако более эффективные методы отсутствуют. Также в литературе не описаны методы решения задач, в которых присутствуют поверхностные токи.